The Dissertation Committee for Eshan Chattopadhyay
certifies that this is the approved version of the following dissertation:

Explicit Two-Source Extractors and More

Committee:

David Zuckerman, Supervisor

Anna Gal

Xin Li

Brent Waters

Explicit Two-Source Extractors and More

by

Eshan Chattopadhyay, B.Tech.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2016

for Mom and Dad

Acknowledgments

I am deeply indebted to my advisor, David Zuckerman for any success I have had in grad school. David tolerated me during my initial days and taught me to be brave in approaching difficult problems. David has been extremely generous with many things over the years: ideas, advice, freedom and support, and I have vastly benefitted both personally and as a researcher. His remarkable optimism and bold ideas to evade seemingly unavoidable barriers led to some of the stronger results in this thesis. It has been a pleasure knowing and working with David, and is clearly one of the best things of my grad life.

Apart from David, I am very thankful to Xin Li, who has had a big influence on my research. Xin has been an amazing collaborator and friend, and I have learnt a lot of things from him. I thank him for hosting me at Johns Hopkins which turned out to be a very productive time. I thank Vipul Goyal for hosting me for a productive and fun-filled summer at Microsoft Research, India. Vipul has a great attitude towards research, and apart from technical stuff, I hope I have absorbed some of his approach towards research. I also thank Satya Lokam and Bhavana Kanukurthi for discussions during my stay at MSR from which I learnt a lot. Many thanks to my thesis committee members: Anna Gal, Xin Li, Brent Waters and David Zuckerman for patiently reading my thesis and useful comments.

I have been lucky to visit other research institutes during my graduate studies and I have greatly benefitted from these trips. I thank Avi Wigderson and Ran Raz for arranging a trip to the Institute of Advanced Studies. I thank Yael Kalai for discussions at various points and inviting me

to Microsoft Research, New England. I thank Yevgeniy Dodis for many valuable discussions and sharing many research directions. I thank the organizers of the Oberwolfach Complexity workshop: Peter Burgisser, Oded Goldreich, Madhu Sudan, and Salil Vadhan for a fun-filled week of talks and discussions on complexity. I also thank the organizers of the workshop on the Foundations of Randomness: Artur Ekert, Renato Renner, Miklos Santha, Umesh Vazirani and Thomas Vidick for a memorable week in South Africa. I thank Assaf Naor for arranging a visit to the Simons Algorithms and Geometry meeting in New York. I also wish to thank Divesh Aggarwal, Noga Alon, Mark Braverman, Gil Cohen, Zeev Dvir, Oded Goldreich, Venkat Guruswami, Robert Kleinberg, Adam Klivans, Or Meir, Raghu Meka, Eric Price, Anup Rao, Alexander Razborov, Omer Reingold, Michael Saks, Igor Shinkar, Thomas Steinke, Avishay Tal, Salil Vadhan, Omri Weinstein, Daniel Wichs and Henry Yuen for useful discussions and enjoyable conversations at various points.

I am very grateful to my friends and office mates: Aayush, Abhishek, Adarsh, Ameya, Ankit Garg, Ankit Rawat, Akshay, Fu, Harsh, John, Kartik, Matteo, Pravesh, Rakesh, Rashish, Rishabh, Saurabh, Sebastien, Siddhesh, Teja, Thodoris, Udit, Venkata, Vishal, Xue (and many others) for the many discussions on random topics, coffee breaks at odd hours and support at various points which helped me remain motivated even when I wasn't making much progress in research.

Finally, I would like to thank my family and people closest to me for their incredible belief in my abilities and invaluable support over these years (despite various hardships), without which this thesis would have been impossible. A special thanks to my parents for raising me right and providing me with a happy childhood, and my grandmothers for spoiling me.

Eshan Chattopadhyay

The University of Texas at Austin

May 2016

Explicit Two-Source Extractors and More

Publication No. _____

Eshan Chattopadhyay, Ph.D.
The University of Texas at Austin, 2016

Supervisor: David Zuckerman

In this thesis we study the problem of extracting almost truly random bits from imperfect sources of randomness. This is motivated by the wide use of randomness in computer science, and the fact that most accessible sources of randomness generate correlated bits, and at best contain some amount of entropy. We follow Chor and Goldreich [CG88] and Zuckerman [Zuc90], and model weak sources using min-entropy, where an (n, k)-source $\mathbf{X}$ is a distribution on n bits and takes any string x with probability at most 2^{-k}. It is known that it is impossible to extract random bits from a single (n, k)-source, and Chor and Goldreich [CG88] raised the question of extracting randomness from two such independent (n, k)-sources. Existentially, such 2-source randomness extractors exist for min-entropy $k \geq \log n + O(1)$, but the best known construction prior to work in this thesis requires min-entropy $k \geq 0.499n$ [Bou05b]. One of the main contributions of this thesis is an explicit 2-source extractor for min-entropy $\log^C n$, for some constant C.

Other results in this thesis include improved ways of extracting random bits from various other sources of randomness, as well as stronger notions of randomness extraction. Our results have applications in privacy amplification [BBR88, Mau92, BBCM95], which is a classical problem in information cryptography, and give protocols that achieve almost optimal parameters. Other applications include explicit constructions of non-malleable codes, which is a relaxation of the notion of error-detection codes and have applications in tamper-resilient cryptography [DPW10].

Contents

Chapter 1

Introduction

In this thesis we study objects known as randomness extractors. Informally, an extractor is a tool to purify a source of randomness. The need for such extractors arises out of the following two reasons. First, randomness is widely used in various areas of computer science, e.g., algorithms, distributed computing, cryptography, stochastic simulations of complex systems and more. Second, most sources of randomness that are easily accessible produce bits that are biased and correlated.

It is very common that randomized algorithms are often much simpler than their (known) deterministic counterparts, and also outperform them. In fact, in many cases such as polynomial identity testing, it is open to find efficient deterministic algorithms for problems with simple randomized algorithms. Here a major open question is if every efficient randomized algorithm has a deterministic counterpart, or more technically whether P = BPP. Further, in cryptography, it is possible to prove that many of the protocols become provably impossible to execute without access to high quality sources of randomness [DOPS04].

However a major problem in practice is the lack of good quality sources of randomness. For example, a common way operating systems collect random bits (such as Linux) is to maintain an entropy pool. This entropy pool is typically filled by a device known as a hardware random number generator (HRNG) that uses some physical phenomenon (e.g., radioactive phenomenon,

Zener diodes, clock drift) for generating randomness. There is also work showing Bitcoin [BCG15] as a potential source of randomness. However, in most of these sources the bits produced often follow certain patterns and at best only contain some amount of entropy. In practice, to derive uniform bits, often a cryptographic hash function is applied to this imperfect source of randomness and used in applications. However, there is no theoretical guarantee that the bits produced are actually uniformly random, which can be a major issue, for example, if these bits are being used to carry out important cryptographic protocols. Another motivation to study weak sources of randomness comes from cryptography. For example, consider a shared key $\mathbf{S}$ that is uniform on n bits and is being used for executing some cryptographic protocol. An adversary who gains partial information about this secret can be modeled as a function $f : \{0,1\}^n \to \{0,1\}^m$, $m < n$, with the partial information being the random variable $\mathbf{Y} = f(\mathbf{S})$. Thus, the secret $\mathbf{S}$ conditioned on the leak $\mathbf{Y}$ is now only weakly random. These applications motivate the need to construct functions (extractors) that provably output uniform bits given access to such weak sources of randomness.

1.0.1 A Brief History of Extractors

The first work of modeling a weak source dates back to the 1950's when von Neumann [vN51] considered extracting random bits from a stream of independent bits with the same unknown bias. This was considerably generalized by Blum [Blu86], who designed an explicit extractor for sources generated by a constant sized Markov chain. Santha and Vazirani [SV86] further generalized this model and introduced SV-sources (in [SV86], these sources are called as slightly-random sources), where each bit in the sequence is "slightly random" and takes the value 0 with probability in the range $(\delta, 1-\delta)$, $0 < \delta < 1/2$, for any conditioning of the previous bits in the sequence. They proved that it is impossible to extract from a single SV-source and gave an efficient algorithm to extract from $O(\log n \log^* n)$ independent SV-sources, and left as an open problem to extract from two independent SV-sources.

Chor and Goldreich [CG88] introduced a model of weak sources called block sources, and Zuckerman [Zuc90] generalized this to model weak sources using the notion of min-entropy. A

source $\mathbf{X}$ on n bits is said to have min-entropy at least k if for any x, $\mathbf{Pr}[\mathbf{X} = x] \leq 2^{-k}$.

Definition 1.0.1. *The min-entropy of a source* $\mathbf{X}$ *is defined to be:* $H_\infty(\mathbf{X}) = \min_x(-\log(\Pr[\mathbf{X} = x]))$. *The min-entropy rate of a source* $\mathbf{X}$ *on* $\{0,1\}^n$ *is defined to be* $H_\infty(\mathbf{X})/n$. *Any source* $\mathbf{X}$ *on* $\{0,1\}^n$ *with min-entropy at least* k *is called an* (n,k)*-source.*

This is now the standard way of modeling a weak source. However, it turns out that the class of (n,k)-sources is too general and the following simple lemma shows that it is impossible to extract from this class. We first introduce the notion of an extractor for a class of sources.

We use statistical (or variation) distance to measure the performance of an extractor in terms of the closeness of the output to the uniform distribution.

Definition 1.0.2. *The statistical distance between two distributions* $\mathcal{D}_1$ *and* $\mathcal{D}_2$ *over some universal set* Ω *is defined as* $|\mathcal{D}_1 - \mathcal{D}_2| = \frac{1}{2}\sum_{d\in\Omega} |\mathbf{Pr}[\mathcal{D}_1 = d] - \mathbf{Pr}[\mathcal{D}_2 = d]|$. *We say* $\mathcal{D}_1$ *is* ϵ*-close to* $\mathcal{D}_2$ *if* $|\mathcal{D}_1 - \mathcal{D}_2| \leq \epsilon$ *and denote it by* $\mathcal{D}_1 \approx_\epsilon \mathcal{D}_2$.

Definition 1.0.3. *An efficiently computable function* $\mathrm{Ext} : \{0,1\}^n \to \{0,1\}^m$ *is an (deterministic) extractor for a class of sources* $\mathcal{X}$ *with error* ϵ *if, for any source* $\mathbf{X} \in \mathcal{X}$, $|f(\mathbf{X}) - \mathbf{U}_m| \leq \epsilon$.

Lemma 1.0.4 ([CG88])**.** *There cannot exist an extractor for the class of* $(n, n-1)$*-sources with error* $< 1/2$.

Proof. Suppose $\mathrm{Ext} : \{0,1\}^n \to \{0,1\}$ is an extractor that extracts for min-entropy $n-1$ with error $\epsilon < 1/2$. Since $|\mathrm{Ext}^{-1}(0)| + |\mathrm{Ext}^{-1}(1)| = 2^n$, W.lo.g, let $|\mathrm{Ext}^{-1}(0)| \geq 2^{n-1}$. Let $\mathbf{X}$ be a source uniform on the set $\mathrm{Ext}^{-1}(0)$. Clearly, the min-entropy of $\mathbf{X}$ is at least $n-1$ but Ext is constant on $\mathbf{X}$, which contradicts the assumption that the error of Ext is less than $1/2$. □

To circumvent this difficulty, Chor and Goldreich suggested the problem of extracting from two or more independent sources.

Definition 1.0.5 (2-source extractor)**.** *A function* $\text{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ *is called a* (k,ϵ)*-two-source extractor if for independent* (n,k)*-sources* $\mathbf{X}$ *and* $\mathbf{Y}$*, we have*

$$|\text{Ext}(\mathbf{X},\mathbf{Y}) - \mathbf{U}_m| \leq \epsilon.$$

Ext *is said to be strong in* $\mathbf{Y}$ *if it also satisfies* $|(\text{Ext}(\mathbf{X},\mathbf{Y}),\mathbf{Y}) - (\mathbf{U}_m,\mathbf{Y})| \leq \epsilon$*, where* $\mathbf{U}_m$ *is independent from* $\mathbf{Y}$*.*

A simple probabilistic argument shows the existence of 2-source extractors for min-entropy $k \geq \log n + 2\log(1/\epsilon) + 1$. However, for applications, one is interested in efficiently constructing such extractors. Chor and Goldreich [CG88] used Lindsey's Lemma to show that the inner-product function (see Theorem 2.5.3) is a 2-source extractor for min-entropy more than $n/2$. No further progress was made for around 20 years, when Bourgain [Bou05b] constructed a 2-source extractor for min-entropy $0.499n$. His result was based on breakthroughs in the area of additive combinatorics. Raz [Raz05] obtained an improvement in terms of total min-entropy, and constructed 2-source extractors requiring one source with min-entropy more than $n/2$ and the other source with min-entropy $O(\log n)$. There is also a different 2-source extractor on the Paley graph function matching the entropy bounds of [Raz05] (see Theorem 2.5.4). Prior to work in this thesis, it was a challenging open problem to construct a 2-source extractor that works when both sources have min-entropy significantly smaller than $n/2$.

An explicit 2-source extractor directly yields explicit Ramsey graphs, a central object in extremal combinatorics. Recall that a graph on N vertices is called a K-Ramsey graph if it does not contain any independent set or clique of size K. In 1947, it was shown by Erdös in one of the first applications of the probabilistic method that there exists K-Ramsey graphs for $K = 2\log N$. He posed as a challenge to explicitly construct such a graph, and this has drawn a lot of attention over the last 69 years. Frankl and Wilson [FW81] used intersection theorems to construct K-Ramsey graphs on N vertices, with $K = 2^{O(\sqrt{\log N \log\log N})}$. This remained the best known construction for a long time, with many other constructions [Alo98, Gro00, Bar06, Gop14] achieving the same bound.

Finally, subsequent works by Barak et al. [BKS+10, BRSW12] obtained a significant improvement and gave explicit constructions of K-Ramsey graphs, with $K = 2^{2^{\log^{1-\alpha}(\log N)}}$, for some absolute constant α.

An impressive line of work considered the problem of constructing extractors having access to multiple independent sources. Several researchers managed to construct excellent extractors using a constant number of sources [BIW06, Rao09a, RZ08, Li11a, Li13a, Li13b, Li15e, Coh15a], with the best known result being a 3-source extractor construction for $(\log n)^C$ min-entropy by Li [Li15e].

Raz and Yehudayoff [RY11] introduced a natural generalization of the class of independent sources, which called interleaved sources. Roughly, the symbols from C independent source are mixed(in some unknown order) into one long string and given as input to the extractor. Besides being a natural generalization of independent sources, the original motivation for studying these sources came from an application found by Raz and Yehudayoff [RY11] in proving lower bounds for arithmetic circuits. Further, such extractors give examples of explicit functions with high best-partition communication complexity. Using the probabilistic method, one can show that extractors exist for $C = 2$ and $k = \Omega(\log n)$. The construction in [RY11] works however works for $k > (1-\delta)n$ and $C = 2$, where δ is a small constant.

A different line of work considered the problem of simulating randomized algorithms with access to only weak sources of randomness [VV85, CG88, Zuc96, SSZ95, ACRT97]. This led to the introduction of the notion of seeded extractors [NZ96]. Informally, a seeded extractor uses a short uniform seed to extract randomness out of an (n, k)-source $\mathbf{X}$ (see Chapter 2 for a formal definition). A long line of work spanning two decades culminated in excellent constructions of seeded extractors (see [LRVW03, GUV09, DKSS09] for current optimal constructions). Further various applications of seeded extractors were found in seemingly unrelated areas like inapproximabilty [Zuc96, Uma99, MU02], error correcting codes [TZ04, Gur04a], expander graphs [WZ93] (see also [NT99] for more applications).

In another line of work, Trevisan and Vadhan [TV00] introduced the problem of constructing seedless extractors for the class of samplable sources, where the weak random source is generated

by a computationally bounded algorithm. The simplest sources in this model are bit-fixing sources. Informally, a bit-fixing source is a source where some subset of the bits are fixed and the remaining ones chosen in some random way. Such sources have applications in exposure resilient cryptography and have been investigated in a line of work [CGH$^+$85, KZ07a, GRS06, Rao09b]. Generalizing oblivious bit-fixing sources (see Chapter 5 for a definition) are a class of sources called as affine sources. Here the source is assumed to be uniform on some unknown affine subspace of $\mathbb{F}_p^n$ of dimension k. For $p = 2$, (which is the most interesting setting in applications to computer science), the best known affine extractor until very recently worked for $k \geq n/\sqrt{\log \log n}$ [Bou07,Li11b,Yeh11] (a recent work of Li [Li15c] improves this to $k \geq \log^C n$ using components from work in this thesis). For larger p, Gabizon and Raz [GR08] constructed almost optimal extractors even for $k = 1$. Ben-Sasson and Zewi [BSZ11] showed some connections between affine extractors and 2-source extractor based on conjectures in additive combinatorics. In [TV00], they constructed explicit extractors for sources generated by polynomial sized circuits based on strong complexity-theoretic assumptions. Kamp, Rao, Vadhan and Zuckerman [KRVZ11] studied the problem of constructing extractors for small-space sources, where the weak source is generated by a small width branching program. Roughly, their extractor construction works for min-entropy $n^{1-\delta}$, for some small absolute constant δ.

Recently, Dodis and Wichs [DW09] initiated the study of seeded non-malleable extractors with applications to cryptography. These extractors strengthen the notion of seeded extractors in a non-trivial way. Very informally, a non-malleable extractor has to satisfy the property that the output of the extractor looks random even to an adversary that has access to the output of the extractor evaluated on a correlated seed. In some more detail, suppose $\mathbf{X}$ is an (n, k)-source, and $\mathcal{A} : \{0,1\}^d \to \{0,1\}^d$ is a function such that $\mathcal{A}(y) \neq y$ for all y. We think of $\mathcal{A}$ as an adversary, and call such functions as tampering functions. Let $\mathbf{Y}$ be a uniform independent seed on d bits. Then, a non-malleable extractor $\mathrm{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ satisfies the property that for most fixings of $\mathbf{Y} = y$, we have $\mathrm{nmExt}(\mathbf{X}, y), \mathrm{nmExt}(\mathbf{X}, \mathcal{A}(y)) \approx \mathbf{U}_m, \mathrm{nmExt}(\mathbf{X}, \mathcal{A}(y))$. It turns out that this property is quite non-trivial to satisfy, and the first explicit construction was found by Dodis,

Li, Wichs and Zuckerman [DLWZ14]. Subsequent works [CRS14, Li12a, Li12b, DY13] improved various parameters, but all these constructions required the min-entropy of the source $\mathbf{X}$ to be at least $0.499n$. However, existentially the work [DW09] proved the existence of such extractors for $k \geq \log n$ (for polynomially small error).

The main applications of such non-malleable extractors comes from the problem of privacy amplification with an active adversary [BBR88, Mau92, BBCM95]. As a basic problem in information theoretic cryptography, privacy amplification deals with the case where two parties want to communicate with each other to convert their shared secret weak random source $\mathbf{X}$ into shared secret nearly uniform random bits. On the other hand, the communication channel is watched by an adversary Eve, who has unlimited computational power. To make this task possible, we assume two parties have local (non-shared) uniform random bits. If Eve is passive (i.e., can only see the messages but cannot change them), this problem can be solved easily by using strong seeded extractors. However, in the case where Eve is active (i.e., can change, delete and reorder messages), the problem becomes much more complicated. The major challenge here is to design a protocol that uses as few interactions as possible, and outputs a uniform random string $\mathbf{R}$ that has length as close to $H_\infty(\mathbf{X})$ as possible (the difference is called the *entropy loss*).

A 2-source variant of non-malleable extractors, called seedless non-malleable extractors, was introduced by Cheraghchi and Guruswami [CG14b]. Here, roughly, the tampering functions act on both $\mathbf{X}$ and $\mathbf{Y}$. Thus, the guarantee we would want is $\mathrm{nmExt}(\mathbf{X}, \mathbf{Y}), \mathrm{nmExt}(f(\mathbf{X}), g(\mathbf{Y})) \approx \mathbf{U}_m, \mathrm{nmExt}(f(\mathbf{X}), g(\mathbf{Y}))$, where $\mathbf{X}$ and $\mathbf{Y}$ are independent weak sources and f, g are arbitrary tampering functions (with one of them not mapping any input to itself). Their main motivation for initiating the study of these objects are in applications to non-malleable codes. Non-malleable codes (introduced by Dziembowski, Pietrzak and Wichs [DPW10]) are a natural weakening of error-detecting codes in hope to handle more severe forms of tampering on the codeword (see Section 12 for a definition). These codes also have applications in tamper-resilient cryptography [DPW10]. In [CG14b], they showed a black-box way of constructing non-malleable codes via explicit seedless non-malleable extractors. Further, they showed the existence of such non-malleable extractors for

min-entropy $\Omega(\log n)$. However, no known constructions of such seedless non-malleable extractors were known prior to work in this thesis, and it was posed as an open problem in [CG14b] to construct such an extractor even for full min-entropy (i.e., $k = n$).

1.0.2 Our Results

2-Source Extractors One of the main contributions of this thesis is an explicit 2-source extractor that works for min-entropy $k \geq \log^C n$, for some constant C. This is based on joint work with David Zuckerman [CZ16a]. In subsequent work with Xin Li [CL16a], we improve the constant C. We present this in Chapter 6. The construction needs material that is developed in Chapters 4 and 5.

Ramsey Graphs As a corollary of our 2-source extractor, we obtain explicit K-Ramsey graphs on $N = 2^n$ vertices with $K = 2^{(\log \log N)^C}$ for some constant C. This result was also obtained by an independent work by Cohen [Coh16c], who constructed a weaker object called a 2-source disperser (see Definition 2.5.2) for min-entropy $\log^C n$ to obtain this result.

Seeded Non-Malleable Extractors and Privacy Amplification We give explicit constructions of seeded non-malleable extractors that requires min-entropy $k = \Omega(\log^2 n/\epsilon)$ and seed-length $d = O(\log^2 n/\epsilon)$, where ϵ is an error parameter. In fact our construction is more general, and this is a crucial ingredient in our 2-source extractor construction. This result is based on joint work with Vipul Goyal and Xin Li [CGL16]. Subsequently, we improve this to $k = \Omega(\log^{1+o(1)} n/\epsilon)$ and seed-length $d = O(\log^{1+o(1)} n/\epsilon)$. This is based on joint work with Xin Li [CL16a]. This improvement is crucial to obtain new results in the problem of privacy amplification, where we obtain a protocol with almost optimal parameters (a substantial amount of research over the last 25 years has focussed on obtaining such a protocol, and our result is nearly optimal). We present these results in Chapter 4. A substantial amount of tools for these results are developed in Chapter 3. The techniques in Chapter 3 crucially rely on the powerful technique of "alternating extraction" that was introduced by Dziembowski and Pietrzak [DP07].

Resilient Functions An ingredient in our 2-source extractor construction are functions that have low influence with respect to small subsets of co-ordinates. Such functions were initially studied by Ben-Or and Linial [BL85] when they introduced the perfect information model. We obtain new results on explicitly constructing such resilient functions and present this in Chapter 5. This is based on joint work with David Zuckerman [CZ16a].

Small-Space Sources We give improved extractors for small space sources that work for min-entropy $k = n^{o(1)}$. This is based on joint work with Xin Li [CL16b]. The results are presented in Chapter 9. This uses results from Chapter 8.

Sumset Sources We generalize the class of affine sources and study sources of the form $\mathbf{X}_1 + \ldots + \mathbf{X}_C$ where each $\mathbf{X}_i$ is an independent source on $\mathbb{F}_2^n$ (the addition being the usual vector addition). We show how to extract when each $\mathbf{X}_i$ has min-entropy at least $\log^C n$. We also show applications of sumset extractors to extract from many other weak sources that have been previously studied. This is based on joint work with Xin Li [CL16b]. We present the results in Chapter 8. We use some components from Chapter 3 for this construction.

Seedless Non-Malleable Extractors and Non-Malleable codes We present two constructions of seedless non-malleable extractors. The first construction uses 10 sources (instead of 2, generalizing the definition so that each source is tampered) and is based on joint work with David Zuckerman [CZ14]. As a result, we obtain the first construction of a non-malleable code with constant rate in a well studied "split-state" model. The second construction uses just 2 sources and thus resolves the open question of [CG14b]. Further, this construction generalizes to handle multiple tamperings, and using this we give the first explicit constructions of non-malleable codes that can handle multiple attacks in the information theoretic setting. We present our result on non-malleable extractors in Chaper 11 and our results on non-malleable codes in Chapter 12. Some of the results uses components from Chapter 3.

Interleaved Sources We give improved constructions of extractors for interleaved sources. We use Chapter 10 to present these results. The results are based on joint works with Xin Li and David Zuckerman [CZ16b, CL16b].

Multi-Source Extractors The best known multi-source extractor (in terms of min-entropy) is from a recent work of Cohen and Schulman [CS16] and requires $O(1/\delta)+O(1)$ independent sources, each with min-entropy at least $\log^{1+\delta}(n)$. We improve this result and give extractors that work for $O(1)$ (an absolute constant) independent sources, each with min-entropy $\log^{1+o(1)} n$. This is based on joint work with Xin Li [CL16a].

Chapter 2

Preliminaries

We use $\mathbf{U}_m$ to denote the uniform distribution on $\{0,1\}^m$, and $\mathbf{U}_S$ to denote the uniform distribution on any set S.

For any integer $t > 0$, $[t]$ denotes the set $\{1, \ldots, t\}$.

We use bold capital letters for random variables and samples as the corresponding small letter, e.g., $\mathbf{X}$ is a random variable, with x being a sample of $\mathbf{X}$.

For an $\ell \times m$ matrix V, and any $S \subseteq \ell$, $|S| = q$, we use V_S to denote the $q \times m$ sub-matrix of V corresponding to the rows indexed by S. If $S = \{i\}$ is a singleton, we use V_i instead of $V_{\{i\}}$.

A distribution $\mathcal{D}$ on n bits is t-wise independent if the restriction of $\mathcal{D}$ to any t bits is uniform. Further $\mathcal{D}$ is (t, ϵ)-wise independent if the distribution obtained by restricting $\mathcal{D}$ to any t coordinates is ϵ-close to uniform.

For any integer $M > 0$, let $e_M(x) = e^{\frac{2\pi i x}{M}}$.

2.1 Seeded Extractors

Definition 2.1.1. *A function* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is a* (k, ϵ)*-seeded extractor if for any source* $\mathbf{X}$ *of min-entropy* k, $|\mathrm{Ext}(\mathbf{X}, \mathbf{U}_d) - \mathbf{U}_m| \leq \epsilon$. Ext *is called a strong seeded extractor if* $|(\mathrm{Ext}(\mathbf{X}, \mathbf{U}_d), \mathbf{U}_d) - (\mathbf{U}_m, \mathbf{U}_d)| \leq \epsilon$, *where* $\mathbf{U}_m$ *and* $\mathbf{U}_d$ *are independent. Further, if for each* $s \in \mathbf{U}_d$,

$\mathrm{Ext}(\cdot, s) : \{0,1\}^n \to \{0,1\}^m$ *is a linear function, then* Ext *is called a linear seeded extractor.*

We recall an explicit seeded extractor construction with almost optimal parameters.

Theorem 2.1.2 ([GUV09])**.** *For any constant $\alpha > 0$, and all integers $n, k > 0$ there exists a polynomial time computable strong-seeded extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with* $d = O(\log n + \log(1/\epsilon))$ *and* $m = (1-\alpha)k$*.*

For some applications we need to ensure that for each $x \in \{0,1\}^n$, $\mathrm{Ext}(x, s_1) \neq \mathrm{Ext}(x, s_2)$ whenever $s_1 \neq s_2$. A simple way to ensure this is to concatenate the seed to the output of Ext, though it is no longer strong. We record this formally.

Corollary 2.1.3. *For any constant $\alpha > 0$, and all integers $n, k > 0$ there exists a polynomial time computable seeded extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with* $d = O(\log n + \log(1/\epsilon))$ *and* $m = (1-\alpha)k$*. Further for all* $x \in \{0,1\}^n$, $\mathrm{Ext}(x, s_1) \neq \mathrm{Ext}(x, s_2)$ *whenever* $s_1 \neq s_2$*.*

We also use the following strong seeded extractor constructed by Zuckerman [Zuc07] that achieves seed length $\log(n) + O(\log(\frac{1}{\epsilon}))$ to extract from any source with constant min-entropy rate.

Theorem 2.1.4 ([Zuc07])**.** *For all $n > 0$ and constants $\alpha, \delta, \epsilon > 0$ there exists an efficient construction of a $(k = \delta n, \epsilon)$-strong seeded extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with* $m \geq (1-\alpha)k$ *and* $D = 2^d = O(n)$*.*

In some of our constructions, we require explicit linear seeded extractors with strong parameters.

Theorem 2.1.5 ([Tre01, RRV02])**.** *For every $n, k, m \in \mathbb{N}$ and $\epsilon > 0$, with $m \leq k \leq n$, there exists an explicit strong linear seeded extractor* $\mathrm{LExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *for min-entropy k and error ϵ, where* $d = O\left(\frac{\log^2(n/\epsilon)}{\log(k/m)}\right)$*.*

A drawback of the above construction is that the seed length is $\omega(\log n)$ for sub-polynomial min-entropy. An improved construction of Li [Li15c] achieves $O(\log n)$ seed length for even poly-logarithmic min-entropy.

Theorem 2.1.6 ([Li15c]). *There exists a constant $c > 1$ such that for every $n, k \in \mathbb{N}$ with $c \log^8 n \leq k \leq n$ and any $\epsilon \geq 1/n^2$, there exists a polynomial time computable linear seeded extractor* $\mathrm{LExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *for min-entropy k and error ϵ, where $d = O(\log n)$ and $m \leq \sqrt{k}$.*

We record useful lemma which shows that seeded extractors work even when the seed is not fully uniform, but has sufficiently large min-entropy.

Lemma 2.1.7. *Let* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a strong seeded extractor for min-entropy k, and error ϵ. Let $\mathbf{X}$ be a (n,k)-source and let $\mathbf{Y}$ be a source on $\{0,1\}^d$ with min-entropy $d - \lambda$. Then,*

$$|\mathrm{Ext}(\mathbf{X}, \mathbf{Y}) \circ \mathbf{Y} - \mathbf{U}_m \circ \mathbf{Y}| \leq 2^{\lambda} \epsilon.$$

Proof. Since Y is a source with min-entropy $d - \lambda$, we can assume it is uniform on a set A of size $2^{d-\lambda}$. Thus

$$\begin{aligned} |\mathrm{Ext}(\mathbf{X}, \mathbf{Y}) \circ \mathbf{Y} - \mathbf{U}_m \circ \mathbf{Y}| &= \frac{1}{2^{d-\lambda}} \sum_{y \in A} |\mathrm{Ext}(\mathbf{X}, y) - \mathbf{U}_m| \\ &\leq \frac{1}{2^{d-\lambda}} \sum_{y \in \{0,1\}^d} |\mathrm{Ext}(\mathbf{X}, y) - \mathbf{U}_m| \\ &\leq \frac{1}{2^{d-\lambda}} 2^d \epsilon = 2^{\lambda} \epsilon \end{aligned}$$

where the last inequality uses the fact that Ext is a strong seeded extractor. □

2.2 Conditional Min-Entropy

Definition 2.2.1. *The average conditional min-entropy of a source $\mathbf{X}$ given a random variable $\mathbf{W}$ is defined as*

$$\widetilde{H}_{\infty}(\mathbf{X}|\mathbf{W}) = -\log\left(\mathbf{E}_{w \sim W}\left[\max_x \Pr[\mathbf{X} = x | \mathbf{W} = w]\right]\right) = -\log\left(\mathbf{E}\left[2^{-H_{\infty}(\mathbf{X}|\mathbf{W}=w)}\right]\right).$$

We recall some results on conditional min-entropy from the work of Dodis et al. [DORS08].

Lemma 2.2.2 ([DORS08]). *For any $\epsilon > 0$,* $\mathbf{Pr}_{w\sim\mathbf{W}}\left[H_\infty(\mathbf{X}|\mathbf{W}=w) \geq \widetilde{H}_\infty(\mathbf{X}|\mathbf{W}) - \log(1/\epsilon)\right] \geq 1-\epsilon$.

Lemma 2.2.3 ([DORS08]). *If a random variable $\mathbf{Y}$ has support of size 2^ℓ, then $\widetilde{H}_\infty(\mathbf{X}|\mathbf{Y}) \geq H_\infty(\mathbf{X}) - \ell$.*

We require extractors that can extract uniform bits when the source only has sufficient conditional min-entropy.

Definition 2.2.4. *A (k,ϵ)-seeded average case seeded extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *for min-entropy k and error ϵ satisfies the following property: For any source $\mathbf{X}$ and any arbitrary random variable $\mathbf{Z}$ with $\tilde{H}_\infty(\mathbf{X}|\mathbf{Z}) \geq k$,*

$$\mathrm{Ext}(\mathbf{X}, \mathbf{U}_d), \mathbf{Z} \approx_\epsilon \mathbf{U}_m, \mathbf{Z}.$$

It was shown in [DORS08] that any seeded extractor is also an average case extractor.

Lemma 2.2.5 ([DORS08]). *For any $\delta > 0$, if* Ext *is a (k,ϵ)-seeded extractor, then it is also a $(k + \log(1/\delta), \epsilon + \delta)$-seeded average case extractor.*

2.3 Some Probability Lemmas

We say that a distribution D_1 is ϵ-close to another distribution D_2 if $|D_1 - D_2| \leq \epsilon$.

Definition 2.3.1. *The collision probability of a distribution D is defined as :* $\mathrm{cp}(D) = \Pr[D = D']$, *where D' is independent and identicaly distributed as D.*

For the sake of convenience, we make the following definition.

Definition 2.3.2. *For a set A, define* $\mathrm{cp}(A)$ *to be the collision probability of the uniform distribution on A.*

The following lemma was proved in [BIW06].

Lemma 2.3.3. *Let D be a distribution with $\mathrm{cp}(D) = \frac{1}{KL}$. Then D is $L^{-1/2}$-close to a distribution with min-entropy at least $\log K$.*

Definition 2.3.4. *We say that a distribution D on a set S is a convex combination of distributions $D_1, \ldots, D_l$ on S if there exists non-negative constants (called weights) $w_1, \ldots, w_l$ with $\sum_{i=1}^{l} w_i = 1$ such that $\Pr[D = s] = \sum_{i=1}^{l} w_i \cdot \Pr[D_i = s]$ for all $s \in S$. We use the notation $D = \sum_{i=1}^{l} w_i \cdot D_i$ to denote the fact that D is a convex combination of the distributions $D_1, \ldots, D_l$ with weights $w_1, \ldots, w_l$.*

Definition 2.3.5. *For random variables X and Y, we use $X|Y$ to denote a random variable with distribution: $\Pr[(X|Y) = x] = \sum_{y \in \mathrm{support}(Y)} \Pr[Y = y] \cdot \Pr[X = x|Y = y]$.*

We note the following lemma which follows from the above definitions.

Lemma 2.3.6. *Let X and Y be distributions on a set S such that $X = \sum_{i=1}^{l} w_i \cdot X_i$ and $Y = \sum_{i=1}^{l} w_i \cdot Y_i$. Then $|X - Y| \leq \sum_i w_i \cdot |X_i - Y_i|$.*

The following result on min-entropy was proved by Maurer and Wolf [MW97].

Lemma 2.3.7. *Let $\mathbf{X}, \mathbf{Y}$ be random variables such that $\mathbf{Y}$ takes at ℓ values. Then*

$$\mathbf{Pr}_{y \sim \mathbf{Y}} \left[H_\infty(\mathbf{X}|\mathbf{Y} = y) \geq H_\infty(\mathbf{X}) - \log \ell - \log \left(\frac{1}{\epsilon} \right) \right] > 1 - \epsilon.$$

Lemma 2.3.8 ([BIW06])**.** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_\ell$ be independent random variables on $\{0,1\}^m$ such that $|\mathbf{X}_i - \mathbf{U}_m| \leq \epsilon$. Then, $|\sum_{i=1}^{\ell} \mathbf{X}_i - \mathbf{U}_m| \leq \epsilon^\ell$.*

Lemma 2.3.9 ([GRS06])**.** *Let $\mathbf{X}$ be a random variable taking values in a set S, and let $\mathbf{Y}$ be a random variable on $\{0,1\}^t$. Assume that $|(\mathbf{X}, \mathbf{Y}) - (\mathbf{X}, \mathbf{U}_t)| \leq \epsilon$. Then for every $y \in \{0,1\}^t$,*

$$|(\mathbf{X}|Y = y) - \mathbf{X}| \leq 2^{t+1}\epsilon.$$

Lemma 2.3.10 ([Sha08])**.** *Let $\mathbf{X}_1, \mathbf{Y}_1$ be random variables taking values in a set S_1, and let $\mathbf{X}_2, \mathbf{Y}_2$ be random variables taking values in a set S_2. Suppose that*

1. $|\mathbf{X}_2 - \mathbf{Y}_2| \leq \epsilon_2$.

2. *For every* $s_2 \in S_2$, $|(\mathbf{X}_1|\mathbf{X}_2 = s_2) - (\mathbf{Y}_1|Y_2 = s_2)| \leq \epsilon_1$.

Then

$$|(\mathbf{X}_1, \mathbf{X}_2) - (\mathbf{Y}_1, \mathbf{Y}_2)| \leq \epsilon_1 + \epsilon_2.$$

Using the above results, we record a useful lemma.

Lemma 2.3.11. *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_t$ *be random variables, such that each* $\mathbf{X}_i$ *takes values* 0 *and* 1. *Further suppose that for any subset* $S = \{s_1, \ldots, s_r\} \subseteq [t]$,

$$(\mathbf{X}_{s_1}, \mathbf{X}_{s_2} \ldots, \mathbf{X}_{s_r}) \approx_\epsilon (\mathbf{U}_1, \mathbf{X}_{s_2} \ldots, \mathbf{X}_{s_r}).$$

Then

$$(\mathbf{X}_1, \ldots, \mathbf{X}_t) \approx_{5t\epsilon} \mathbf{U}_t.$$

Proof. We prove this by induction on t. The base case when $t = 1$ is direct. Thus, suppose $t \geq 2$. It follows that

$$(\mathbf{X}_t, \mathbf{X}_1, \ldots, \mathbf{X}_{t-1}) \approx_\epsilon (\mathbf{U}_1, \mathbf{X}_1, \ldots, \mathbf{X}_{t-1}).$$

By an application of Lemma 2.3.9, for any value of the bit b,

$$|(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}|\mathbf{X}_t = b) - (\mathbf{X}_1, \ldots, \mathbf{X}_{t-1})| \leq 4\epsilon.$$

Further, by the induction hypothesis, we have

$$|(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}) - \mathbf{U}_{t-1}| \leq 5(t-1)\epsilon.$$

Thus, by the triangle inequality for statistical distance, it follows that for any value of the bit b,

$$|(\mathbf{X}_1, \ldots, \mathbf{X}_{t-1}|\mathbf{X}_t = b) - \mathbf{U}_{t-1}| \leq (5t-1)\epsilon.$$

Using Lemma 2.3.10 and the fact that $|\mathbf{X}_t - \mathbf{U}_1| \leq \epsilon$, it follows that

$$|(\mathbf{X}_1, \ldots, \mathbf{X}_t) - \mathbf{U}_t| \leq (5t-1)\epsilon + \epsilon = 5t\epsilon.$$

This completes the induction, and the lemma follows. □

We also record the following simple lemma.

Lemma 2.3.12. *Let $\mathbf{X}$ be a source on $\mathbb{F}_p^n$ with min-entropy k. Let $V = \{v_1, \ldots, v_n\}$ be a collection of vectors such that $\dim(\mathrm{span}\{V\}) \geq n - A$. Then $\mathbf{X}_V = \sum_i x_i v_i : x \sim \mathbf{X}$ is a source with min-entropy $\geq k - A \log p$.*

2.4 Sampling Using Weak Sources

A well known way of sampling using weak sources uses randomness extractors. We first introduce a graph-theoretic view of extractors. Any seeded extractor $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ can also be viewed as an unbalanced bipartite graph G_{Ext} with 2^n left vertices (each of degree 2^d) and 2^m right vertices. We use $\mathcal{N}(x)$ to denote the set of neighbours of x in G_{Ext}. We call G_{Ext} the graph corresponding to Ext.

Theorem 2.4.1 ([Zuc97])**.** *Let $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ be a seeded extractor for min-entropy k and error ϵ. Let $D = 2^d$. Then for any set $R \subseteq \{0,1\}^m$,*

$$|\{x \in \{0,1\}^n : ||\mathcal{N}(x) \cap R| - \mu_R D| > \epsilon D\}| < 2^k,$$

where $\mu_R = |R|/2^m$.

Theorem 2.4.2 ([Zuc97])**.** *Let $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ be a seeded extractor for min-entropy k and error ϵ. Let $\{0,1\}^d = \{r_1, \ldots, r_D\}$, $D = 2^d$. Define $\mathrm{Samp}(x) = \{\mathrm{Ext}(x, r_1), \ldots, \mathrm{Ext}(x, r_D)\}$.*

Let $\mathbf{X}$ *be an* $(n, k+k')$*-source. Then for any set* $R \subseteq \{0,1\}^m$*,*

$$\mathbf{Pr}_{x\sim\mathbf{X}}[||\mathrm{Samp}(\mathbf{x}) \cap R| - \mu_R D| > \epsilon D] < 2^{-k'},$$

where $\mu_R = |R|/2^m$*.*

2.5 2-Source Extractors

Definition 2.5.1 (2-source extractor)**.** *A function* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ *is called a* (k,ϵ)*-two-source extractor if for any independent* (n,k)*-sources* $\mathbf{X}$ *and* $\mathbf{Y}$*, we have*

$$|\mathrm{Ext}(\mathbf{X},\mathbf{Y}) - \mathbf{U}_m| \leq \epsilon.$$

Ext *is said to be strong in* $\mathbf{Y}$ *if it also satisfies* $|(\mathrm{Ext}(\mathbf{X},\mathbf{Y}),\mathbf{Y}) - (\mathbf{U}_m,\mathbf{Y})| \leq \epsilon$*, where* $\mathbf{U}_m$ *is independent from* $\mathbf{Y}$*.*

Definition 2.5.2 (2-source Disperser)**.** *A function* $\mathrm{Disp} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ *is called a* (k,ϵ)*-two-disperser if for any independent* (n,k)*-sources* $\mathbf{X}$ *and* $\mathbf{Y}$*,*

$$\mathrm{support}\{\mathrm{Ext}(\mathbf{X},\mathbf{Y})\} = \{0,1\}.$$

We recall a construction of a two-source extractors based on the inner product function [CG88, Zuc91]. This essentially is a stronger version of Lindsey's Lemma. We include a proof for completeness.

Theorem 2.5.3 ([CG88, Zuc91, Rao07])**.** *For all* $m, r > 0$*, with* $q = 2^m, n = rm$*, let* $\mathbf{X}, \mathbf{Y}$ *be independent sources on* $\mathbb{F}_q^r$ *with min-entropy* k_1, k_2 *respectively. Let* IP *be the inner product function over the field* $\mathbb{F}_q$*. Then, we have:*

$$|\mathrm{IP}(\mathbf{X},\mathbf{Y}), \mathbf{X} - U_m, \mathbf{X}| \leq \epsilon, \quad |\mathrm{IP}(\mathbf{X},\mathbf{Y}), Y - \mathbf{U}_m, \mathbf{Y}| \leq \epsilon$$

where $\epsilon = 2^{\frac{-(k_1+k_2-n-m)}{2}}$.

Proof. Let $\mathbf{X}, \mathbf{Y}$ be uniform on sets $A, B \subseteq \mathbb{F}_q^r$ respectively, with $|A| = 2^{k_1}$ and $|B| = 2^{k_2}$. Let ψ be any non-trivial additive character of the finite field $\mathbb{F}_q$. For short, we use $\cdot$ to denote the standard inner product over $\mathbb{F}_q$. We have

$$\begin{aligned}\sum_{y\in B} |\sum_{x\in A} \psi(x\cdot y)| &\leq (|B|)^{\frac{1}{2}} \left(\sum_{y\in\mathbb{F}_q^r} \sum_{x,x'\in A} \psi((x-x')\cdot y)\right)^{\frac{1}{2}} \\ &\leq |B|^{\frac{1}{2}} \left(\sum_{x,x'\in A} \left(\sum_{y\in\mathbb{F}_q^r} \psi((x-x')\cdot y)\right)\right)^{\frac{1}{2}}\end{aligned}$$

where the first inequality follows by an application of the Cauchy-Schwartz inequality. Further, whenever $x \neq x'$, we have

$$\sum_{y\in\mathbb{F}_q^r} \psi((x-x')\cdot y) = 0.$$

Thus, continuing with our estimate, we have

$$\sum_{y\in B} |\sum_{x\in A} \psi(x\cdot y)| \leq |B|^{\frac{1}{2}} \left(|A|q^r\right)^{\frac{1}{2}} = 2^{\frac{n+k_1+k_2}{2}}$$

Thus,

$$E_{\mathbf{Y}} |\mathbf{E}_{\mathbf{X}} \psi(\mathrm{IP}(\mathbf{X},\mathbf{Y}))| \leq 2^{\frac{n-k_1-k_2}{2}}$$

Using Lemma 2.6.1, it now follows that

$$|\mathrm{IP}(\mathbf{X},\mathbf{Y}), \mathbf{Y} - \mathbf{U}_m, \mathbf{Y}| \leq 2^{\frac{n+m-k_1-k_2}{2}}$$

It can be similarly shown that $|\mathrm{IP}(X,Y), X - U_m, X| \leq 2^{\frac{n+m-k_1-k_2}{2}}$. □

The following folklore result on two-source extractors is based on the Paley graph function. The following double character sum estimate was obtained by Karatsuba [Kar71, Kar91].

Theorem 2.5.4 ([Kar71, Kar91])**.** *Let p be any prime. Let χ be a non-trivial multiplicative character of $\mathbb{F}_{p^n}^*$. For any subsets $A, B \subseteq \mathbb{F}_{p^n}$, the following holds: For any integer $\lambda > 0$,*

$$\sum_{a \in A} \left| \sum_{b \in B} \chi(a+b) \right| \leq 2\lambda |A|^{\frac{2\lambda-1}{2\lambda}} (|B| p^{\frac{n}{4\lambda}} + |B|^{\frac{1}{2}} p^{\frac{n}{2\lambda}}).$$

The above theorem can be equivalently restated as a result on 2-source extractors.

Theorem 2.5.5. *Let p be any prime. Let χ be a non-trivial multiplicative character of $\mathbb{F}_{p^n}^*$. For any $\delta > 0$ and independent sources $\mathbf{X}, \mathbf{Y}$ on $\mathbb{F}_{p^n}$ with min-entropy k_1, k_2 respectively, satisfying $k_1 \geq \left(\frac{1}{2} + 3\delta\right) n \log p$ and $k_2 \geq (4 \log n \log p)/\delta$, we have*

$$\mathbf{E}_{x \sim \mathbf{X}} |\mathbf{E}_{y \sim \mathbf{Y}} [\chi(x+y)]| \leq 2^{-\delta k_2}.$$

Proof. Let $\mathbf{X}, \mathbf{Y}$ be flat sources on sets A and B respectively. Thus $|A| = 2^{k_1}$ and $|B| = 2^{k_2}$. Setting $\lambda = \frac{n \log p}{\delta k_2}$ in Theorem 2.5.4 (so that $|B| = 2^{k_2} = p^{\frac{n}{\lambda}}$), we have

$$\begin{aligned}
\mathbf{E}_{x \sim \mathbf{X}} |\mathbf{E}_{y \sim \mathbf{Y}} [\chi(x+y)]| &\leq 2\lambda |A|^{-\frac{1}{2\lambda}} (p^{\frac{n}{4\lambda}} + |B|^{-\frac{1}{2}} p^{\frac{n}{2\lambda}}) \\
&\leq 2\lambda |A|^{-\frac{1}{2\lambda}} (p^{\frac{n}{4\lambda}} + 1) \\
&< 3np^{-\frac{3\delta n}{2\lambda}} \\
&= 2^{\log(3n) - \frac{3k_2 \delta n \log p}{2n \log p}} < 2^{-\delta k_2}.
\end{aligned}$$

□

2.6 Abelian XOR Lemmas

The following lemma is known as Vazirani's XOR Lemma.

Lemma 2.6.1. *Let D be a distribution over $\mathbf{Z}_M$ such that for every nontrivial additive character*

ψ of $\mathbf{Z}_M$, we have $|\mathbf{E}[\psi(D)]| \leq \epsilon$. Then, we have

$$|D - U_M| \leq \epsilon\sqrt{M}.$$

Let $\sigma_M : \mathbf{Z}_N \to \mathbf{Z}_M$ be defined as $\sigma_M(x) = x \pmod{M}$. The following general version of the above XOR lemma was proved in [Rao07].

Lemma 2.6.2 ([Rao07])**.** *Let D be a distribution over $\mathbf{Z}_N$ such that for every non-trivial additive character ψ of $\mathbf{Z}_N$, we have $|\mathbf{E}[\psi(D)]| \leq \epsilon$. Then, for any $M < N$, we have*

$$|\sigma_M(D) - U_M| \leq O(\epsilon \log N\sqrt{M}) + O(M/N).$$

We also record a more generalized form of the XOR Lemma [DLWZ14].

Lemma 2.6.3 ([DLWZ14])**.** *Let D_1, D_2 be distributions over $\mathbf{Z}_N$ such that for arbitrary characters ψ, ϕ of $\mathbf{Z}_N$, we have $|\mathbf{E}[\psi(D_1)\phi(D_2)]| \leq \epsilon$, whenever ψ is nontrivial. Then, for any $M < N$, we have*

$$|(\sigma_M(D_1), \sigma_M(D_2)) - (U_M, \sigma_M(D_2))| = O(\epsilon(\log N)^2 M) + O(M/N).$$

2.7 Finding Primitive Elements in Finite fields

In some of our constructions, we need access to primitive elements in finite fields. There is no known deterministic polynomial time algorithm to find any primitive element of a finite field $\mathbb{F}_{p^n}$. However, there are efficient algorithms known for a weaker task, where the algorithm is only required to output a small set of elements with the guarantee that one of the elements is primitive. The following result is due to Shoup [Sho90].

Theorem 2.7.1 ([Sho90])**.** *Let $p > 0$ be any prime. For all $n > 0$, there exists a deterministic procedure which takes as input n, runs in time* poly(n)*, and outputs a set $S = \{a_1, \ldots, a_l\}$, $l =$* poly(n)*, such that S contains a primitive element of $\mathbb{F}_{p^n}$.*

Chapter 3

Alternating Extraction and its Applications to Breaking Correlations

[1] On a very high level, many of the results in this thesis use explicit objects that break correlations between random variables. In many of the scenarios we consider, the generic setting is the following: Let $\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t$ be correlated random variables, and let $\mathbf{Y}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t$ be random variables independent of $\{\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t\}$. The goal is to construct a function f such that

$$f(\mathbf{X}, \mathbf{Y}), f(\mathbf{X}^1, \mathbf{Y}^1), \ldots, f(\mathbf{X}^t, \mathbf{Y}^t) \approx \mathbf{U}_m, f(\mathbf{X}^1, \mathbf{Y}^1), \ldots, f(\mathbf{X}^t, \mathbf{Y}^t).$$

Clearly this setting is too general, and such an f does not exist. The various objects that we construct make more assumptions on the correlated random variables, and typically $\mathbf{Y}$ is either an independent seed or a weak source with enough min-entropy. In some of the constructions even this is not enough, and we assume access to some kind of 'advice'.

Before we present our actual constructions, we first discuss a toy example to show that seeded extractors are very useful tools in solving problems of this flavour. Let $\mathbf{X}, \mathbf{X}'$ be correlated r.v's, each on n bits, such that $\widetilde{H}_\infty(\mathbf{X}|\mathbf{X}') \geq k$. In such a setting, it is easy to break the correlations

[1] parts of this chapter have been previously published [CGL16, CL16b, CL16a]

between $\mathbf{X}$ and $\mathbf{X}'$ using an independent uniform seed $\mathbf{Y}$. Let $\mathbf{Y}'$ be the tampered version of $\mathbf{Y}$. We use any $(k - \log(1/\epsilon), \epsilon)$-strong seeded extractor Ext and define $\mathbf{Z} = \text{Ext}(\mathbf{X}, \mathbf{Y})$ (and similarly, let $\mathbf{Z}' = \text{Ext}(\mathbf{X}', \mathbf{Y}')$). We prove the correctness of this construction as follows. Fix the r.v $\mathbf{X}'$, and $\mathbf{X}$ still has average conditional min-entropy at least k. Thus, $\text{Ext}(\mathbf{X}, \mathbf{Y})$ is 2ϵ-close to uniform after this fixing, and we can also fix $\mathbf{Y}$ since Ext is a strong extractor. Thus, at this point $\mathbf{Z}'$ is a deterministic function of $\mathbf{Y}'$, and $\mathbf{Z}$ is a deterministic function of $\mathbf{X}$. Thus, we can fix $\mathbf{Z}'$ without affecting the distribution of $\mathbf{Z}$, and hence $f = \text{Ext}$ is a valid construction in this case.

However, in most applications we generally have much weaker guarantees on the correlated r.v's and hence they do not admit such simple solutions. We now introduce the technique of alternating extraction, which informally is a protocol consisting of multiple rounds of extraction using seeded extractors. All our explicit constructions in this chapter are based on some form of the basic alternating extraction method.

The results in this chapter are based on joint works with Vipul Goyal and Xin Li [CGL16, CL16b, CL16a].

3.1 The Basic Alternating Extraction Method and a New Lemma

The method of alternating extraction was introduced by Dziembowski and Pietrzak as a tool to build intrusion resilient secret sharing schemes [DP07]. Subsequently, Dodis and Wichs [DW09] used this method to construct objects known as "look-ahead extractors" and used this to give improved privacy amplification protocols. Since then, this method has an been extremely useful tool in constructing a variety of pseudorandom objects [DW09, Li13a, Li15e, Coh15b, CGL16, Li15c, CL16b, Coh16a, Coh16b].

Alternating Extraction Assume that there are two parties, Quentin with a source $\mathbf{Q}$ and a seed $\mathbf{S}_0$, and Wendy with a source $\mathbf{W}$. The alternating extraction protocol is an interactive process between Quentin and Wendy, and starts off with Quentin sending the seed $\mathbf{S}_0$ to Wendy. Wendy uses $\mathbf{S}_0$ and a strong-seeded extractor Ext_w to extract a seed $\mathbf{R}_0 = \text{Ext}_w(\mathbf{W}, \mathbf{S}_0)$ using $\mathbf{W}$,

and sends $\mathbf{R}_0$ back to Quentin. This constitutes a round of the alternating extraction protocol. In the next round, Quentin uses a strong extractor Ext_q to extract a seed $\mathbf{S}_1 = \mathrm{Ext}_q(\mathbf{Q}, \mathbf{R}_0)$ from $\mathbf{Q}$ using $\mathbf{R}_0$, and sends it to Wendy and so on. The protocol is run for $h+1$ steps, where h is an input parameter. Thus, the following sequence of r.v's is generated: $\mathbf{S}_0, \mathbf{R}_0 = \mathrm{Ext}_w(\mathbf{W}, \mathbf{S}_0), \mathbf{S}_1 = \mathrm{Ext}_q(\mathbf{Q}, \mathbf{R}_0), \ldots, \mathbf{S}_h = \mathrm{Ext}_q(\mathbf{Q}, \mathbf{R}_{h-1}), \mathbf{R}_h = \mathrm{Ext}_w(\mathbf{W}, \mathbf{S}_h)$. Define a look-ahead extractor

$$\mathrm{laExt}(\mathbf{W}, (\mathbf{Q}, \mathbf{S})) = \mathbf{R}_1, \ldots, \mathbf{R}_h.$$

We establish a useful property satisfied by the alternating extraction protocol. This strengthens known results on alternating extraction protocol from previous work [Li13a]. Since stating the result technically involves a lot of parameters, we first informally describe a slightly less general version of the result. Suppose $\mathbf{X}, \mathbf{X}'$ are correlated r.v's, each on n bits, such that $\mathbf{X}$ is an (n,k)-source. Further suppose we have access to r.v's $\mathbf{Y} = (\mathbf{Q}, \mathbf{S}_1), \mathbf{Y}' = (\mathbf{Q}', \mathbf{S}_1')$ with both $\mathbf{Q}, \mathbf{Q}'$ on n bits and both $\mathbf{S}_1, \mathbf{S}_1'$ on d bits, s.t $\{\mathbf{Y}, \mathbf{Y}'\}$ is independent of $\{\mathbf{X}, \mathbf{X}'\}$. Further let $\mathbf{Q}$ be an (n,k)-source. Using $\mathbf{X}$ and $\mathbf{Y}$ in an alternating extraction protocol for h rounds, let the output of the look-ahead extractor be $\mathbf{R}_1, \ldots, \mathbf{R}_h$. Similarly, let $\mathbf{R}_1', \ldots, \mathbf{R}_h'$ be the r.v's output when the alternating extraction protocol is played between $\mathbf{X}'$ and $\mathbf{Y}'$. Then, for any $h < k/10d$, $\mathbf{R}_h$ is close to uniform even conditioned on $\{\mathbf{R}_i : i \in [h-1]\}, \{\mathbf{R}_i' : i \in [h-1]\}$ (with high probability).

We now state and prove this result in full generality. For clarity of presentation, we use the notation: $\mathbf{Z}_{[a,b]}$ to denote the random variable $\mathbf{Z}_a, \ldots, \mathbf{Z}_b$.

Lemma 3.1.1. *Let $\mathbf{X}$ be a (n_w, k_w)-source and let $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)}$ be random variables on $\{0,1\}^{n_w}$ that are arbitrarily correlated with $\mathbf{X}$. Let $\mathbf{Y} = (\mathbf{Q}, \mathbf{S}_1), \mathbf{Y}^{(1)} = (\mathbf{Q}^{(1)}, \mathbf{S}_1^{(1)}), \ldots, \mathbf{Y}^{(t)} = (\mathbf{Q}^{(t)}, \mathbf{S}_1^{(t)})$ be arbitrarily correlated random variables that are independent of $(\mathbf{X}, \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(t)})$. Suppose that $\mathbf{Q}$ is an (n_q, k_q)-source, $\mathbf{S}_1$ is an $(m, m-\lambda)$-source, $\mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(t)}$ are each on n_q bits, and $\mathbf{S}^{(1)}, \ldots, \mathbf{S}^{(t)}$ are each on m bits. Let $\mathrm{Ext}_q, \mathrm{Ext}_w$ be strong seeded extractors that extract m bits at min-entropy k with error ϵ and seed length m. Let* laExt *be the look-ahead extractor for an alternating extraction protocol with parameters u, m, with $\mathrm{Ext}_q, \mathrm{Ext}_w$ being the strong seeded*

extractors used by Quentin and Wendy respectively. Let $\mathrm{laExt}(\mathbf{X},\mathbf{Y}) = \mathbf{R}_1,\ldots,\mathbf{R}_u$ *and for* $j \in [t]$, $\mathrm{laExt}(\mathbf{X}^{(j)},\mathbf{Y}^{(j)}) = \mathbf{R}_1^{(j)},\ldots,\mathbf{R}_u^{(j)}$. *If* $k_w, k_q \geq k + u(t+1)m + 2\log(\frac{1}{\epsilon})$, *then the following holds for each* $i \in [u]$*:*

$$\mathbf{R}_i, \mathbf{R}_{[1,i-1]}, \mathbf{R}^{(1)}_{[1.i-1]}, \ldots, \mathbf{R}^{(t)}_{[1,i-1]}, \mathbf{Q}, \mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(t)} \approx_{\epsilon_i} U_m, \mathbf{R}_{[1,i-1]}, \mathbf{R}^{(1)}_{[1.i-1]}, \ldots, \mathbf{R}^{(t)}_{[1,i-1]}, \mathbf{Q}, \mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(t)}$$

where $\epsilon_i = O(u\epsilon + 2^\lambda \epsilon)$.

Proof. We in fact prove the following stronger claim.

Claim 3.1.2. *For each* $i \in [u]$ *the following hold:*

$$\begin{aligned}&\mathbf{R}_i, \mathbf{R}_{[1,i-1]}, \mathbf{R}^{(1)}_{[1,i-1]}, \ldots, \mathbf{R}^{(t)}_{[1,i-1]}, \mathbf{S}_{[1,i]}, \mathbf{S}^{(1)}_{[1,i]}, \ldots, \mathbf{S}^{(t)}_{[1,i]}, \mathbf{Q}, \mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(t)}\\ \approx_{\epsilon_i}\ & U_m, \mathbf{R}_{[1,i-1]}, \mathbf{R}^{(1)}_{[1,i-1]}, \ldots, \mathbf{R}^{(t)}_{[1,i-1]}, \mathbf{S}_{[1,i]}, \mathbf{S}^{(1)}_{[1,i]}, \ldots, \mathbf{S}^{(t)}_{[1,i]}, \mathbf{Q}, \mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(t)}\end{aligned}$$

and

$$\begin{aligned}&\mathbf{S}_{i+1}, \mathbf{S}_{[1,i]}, \mathbf{S}^{(1)}_{[1,i]}, \ldots, \mathbf{S}^{(t)}_{[1,i]}, \mathbf{R}_{[1,i]}, \mathbf{R}^{(1)}_{[1,i]}, \ldots, \mathbf{R}^{(t)}_{[1,i]}, \mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)}\\ \approx_{\epsilon_i+2\epsilon}\ & U_m, \mathbf{S}_{[1,i]}, \mathbf{S}^{(1)}_{[1,i]}, \ldots, \mathbf{S}^{(t)}_{[1,i]}, \mathbf{R}_{[1,i]}, \mathbf{R}^{(1)}_{[1,i]}, \ldots, \mathbf{R}^{(t)}_{[1,i]}, \mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)}\end{aligned}$$

where $\epsilon_i = 4(i-1)\epsilon + 2^\lambda\epsilon$. *Further, conditioned on* $\mathbf{R}_{[1,i-1]}, \mathbf{R}^{(1)}_{[1,i-1]}, \ldots, \mathbf{R}^{(t)}_{[1,i-1]}, \mathbf{S}_{[1,i]}, \mathbf{S}^{(1)}_{[1,i]}, \ldots, \mathbf{S}^{(t)}_{[1,i]}$, *(a)* $(\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)})$ *is independent from* $(\mathbf{Y}, \mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(t)})$, *(b)* $\mathbf{X}, \mathbf{Q}$ *each have average conditional min-entropy at least* $(u-i)(t+1)m + k + 2\log\left(\frac{1}{\epsilon}\right)$ *and (c)* $\mathbf{R}_i, \mathbf{R}_i^{(1)}, \ldots, \mathbf{R}_i^{(t)}$ *are deterministic functions of* $(\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)})$.

Proof. We prove this claim by induction on i. Let $i = 1$. Since $\mathbf{R}_1 = \mathrm{Ext}_w(\mathbf{X}, \mathbf{S}_1)$, and Ext_w is a strong-seeded extractor, it follows by Lemma 2.1.7 that $\mathrm{Ext}_w(\mathbf{X}, \mathbf{S}_1), \mathbf{S}_1 \approx_{\epsilon_1} U_m, \mathbf{S}_1$, where $\epsilon_1 = 2^\lambda\epsilon$. Thus we can fix $\mathbf{S}_1$, and $\mathbf{R}_1$ is still ϵ_1-close to uniform on average. We note that $\mathbf{R}_1$ is a deterministic function of $\mathbf{X}$. Since the random variables $\mathbf{S}_1^{(1)}, \ldots, \mathbf{S}_1^{(t)}, \mathbf{Q}, \mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(t)}$ are

deterministic functions of $\mathbf{Y}, \mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(t)}$ and thus uncorrelated with $\mathbf{X}$, we have

$$\mathbf{R}_1, \mathbf{S}_1, \mathbf{S}_1^{(1)}, \ldots, \mathbf{S}_1^{(t)}, \mathbf{Q}, \mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(t)} \approx_{\epsilon_1} U_m, \mathbf{S}_1, \mathbf{S}_1^{(1)}, \ldots, \mathbf{S}_1^{(t)}, \mathbf{Q}, \mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(t)}.$$

We fix the random variables $\mathbf{S}_1, \mathbf{S}_1^{(1)}, \ldots, \mathbf{S}_1^{(t)}$. By Lemma 2.2.3, the source $\mathbf{Q}$ has average conditional min-entropy at least $k_q - m(t+1) = k + (u-1)m(t+1) + 2\log\left(\frac{1}{\epsilon}\right)$ after this fixing. Using Lemma 2.2.5 it follows that Ext_q is a $\left(k + \log\left(\frac{1}{\epsilon}\right), 2\epsilon\right)$ strong average case extractor. We also note that $\mathbf{R}_1, \mathbf{R}_1^{(1)}, \ldots, \mathbf{R}_1^{(t)}$ are now deterministic functions of $\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)}$. Thus recalling that $\mathbf{S}_2 = \mathrm{Ext}_q(\mathbf{Q}, \mathbf{R}_1)$, we have $\mathbf{S}_2, \mathbf{R}_1 \approx_{(2\epsilon+\epsilon_1)} U_m, \mathbf{R}_1$, since $\mathbf{R}_1$ is ϵ_1-close to uniform and using the fact that by Lemma 2.2.5 Ext_w is a $\left(k + \log\left(\frac{1}{\epsilon}\right), 2\epsilon\right)$ strong average case extractor. Thus on fixing $\mathbf{R}_1$, $\mathbf{S}_2$ is $(2\epsilon + \epsilon_1)$-close to U_m on average and is a deterministic function of $\mathbf{Y}$. since the random variables $\mathbf{R}_1^{(1)}, \ldots, \mathbf{R}_1^{(t)}$ are deterministic functions of $\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)}$, we thus have

$$\begin{aligned} &\mathbf{S}_2, \mathbf{S}_1, \mathbf{S}_1^{(1)}, \mathbf{S}_1^{(t)}, \mathbf{R}_1, \mathbf{R}_1^{(1)}, \ldots, \mathbf{R}_1^{(t)}, \mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)} \\ \approx_{\epsilon_1+2\epsilon} &\, U_m, \mathbf{S}_1, \mathbf{S}_1^{(1)}, \mathbf{S}_1^{(t)}, \mathbf{R}_1, \mathbf{R}_1^{(1)}, \ldots, \mathbf{R}_1^{(t)}, \mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)} \end{aligned}$$

Further, it still holds that $(\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)})$ is independent from $(\mathbf{Y}, \mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(t)})$. This proves the base case of our induction.

Now suppose that the claim is true for i and we will prove it for $i+1$. Fix the random variables $\mathbf{R}_{[1,i-1]}, \mathbf{R}_{[1,i-1]}^{(1)}, \ldots, \mathbf{R}_{[1,i-1]}^{(t)}, \mathbf{S}_{[1,i]}, \mathbf{S}_{[1,i]}^{(1)}, \ldots, \mathbf{S}_{[1,i]}^{(t)}$. By induction hypothesis, it follows that $\mathbf{X}, \mathbf{Q}$ each have average conditional min-entropy at least $(u-i)m(t+1)+k+2\log\left(\frac{1}{\epsilon}\right)$ after this fixing. We now fix the random variables $\mathbf{R}_i, \mathbf{R}_i^{(1)}, \ldots, \mathbf{R}_i^{(t)}$ (these random variables are deterministic functions of $\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)}$ by induction hypothesis). Thus by Lemma 2.2.3, the source $\mathbf{X}$ has conditional min-entropy at least $(u-i)(t+1)m + k + 2\log\left(\frac{1}{\epsilon}\right) - (t+1)m = (u-i-1)(t+1)m + k + 2\log\left(\frac{1}{\epsilon}\right)$ after this fixing.

Since $\mathbf{S}_{i+1} = \mathrm{Ext}_q(\mathbf{Q}, \mathbf{R}_i)$ is now independent of $\mathbf{X}$ and $(\epsilon_i + 2\epsilon)$-close to U_m on average (by induction hypothesis), it follows that $\mathrm{Ext}_w(\mathbf{X}, \mathbf{S}_{i+1}), \mathbf{S}_{i+1} \approx_{\epsilon_i+4\epsilon} U_m, \mathbf{S}_{i+1}$. Thus on fixing $\mathbf{S}_{i+1}$, the random variable $\mathbf{R}_{i+1} = \mathrm{Ext}_w(\mathbf{X}, \mathbf{S}_{i+1})$ is $(\epsilon_i + 4\epsilon)$-close to U_m on average, and is a deterministic

function of $\mathbf{X}$. We also fix the random variables $\mathbf{S}_{i+1}^{(1)}, \ldots, \mathbf{S}_{i+1}^{(t)}$. Since we have fixed the random variables $\mathbf{R}_i^{(1)}, \ldots, \mathbf{R}_i^{(t)}$, thus $\mathbf{S}_{i+1}^{(1)}, \ldots, \mathbf{S}_{i+1}^{(t)}$ are deterministic functions of $\mathbf{Y}, \mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(t)}$. Hence $\mathbf{R}_{i+1}$ is still ϵ_{i+1}-close to uniform on average and a deterministic function of $\mathbf{X}$ after this fixing. Thus,

$$\begin{aligned}&\mathbf{R}_{i+1}, \mathbf{R}_{[1,i]}, \mathbf{R}_{[1,i]}^{(1)}, \ldots, \mathbf{R}_{[1,i]}^{(t)}, \mathbf{S}_{[1,i+1]}, \mathbf{S}_{[1,i+1]}^{(1)}, \ldots, \mathbf{S}_{[1,i+1]}^{(t)}, \mathbf{Q}, \mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(t)} \\ \approx_{\epsilon_{i+1}} &U_m, \mathbf{R}_{[1,i]}, \mathbf{R}_{[1,i]}^{(1)}, \ldots, \mathbf{R}_{[1,i]}^{(t)}, \mathbf{S}_{[1,i+1]}, \mathbf{S}_{[1,i+1]}^{(1)}, \ldots, \mathbf{S}_{[1,i+1]}^{(t)}, \mathbf{Q}, \mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(t)}.\end{aligned}$$

The source $\mathbf{Q}$ has conditional min-entropy at least $(u-i)(t+1)m+k+2\log\left(\frac{1}{\epsilon}\right)-(t+1)m = (u-i-1)(t+1)m+k+2\log\left(\frac{1}{\epsilon}\right)$.

Recall that $\mathbf{S}_{i+2} = \mathrm{Ext}_q(\mathbf{Q}, \mathbf{R}_{i+1})$. Since Ext_q is a $\left(k+\log\left(\frac{1}{\epsilon}\right), 2\epsilon\right)$ strong average case extractor, it follows that $\mathrm{Ext}_q(\mathbf{Q}, \mathbf{R}_{i+1}), \mathbf{R}_{i+1} \approx_{\epsilon_{i+2}+2\epsilon} U_m$. since the random variables $\mathbf{R}_{i+1}^{(1)}, \ldots, \mathbf{R}_{i+1}^{(t)}$ are deterministic functions of $\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)}$ (recall that we have fixed $\mathbf{S}_{i+1}^{(1)}, \ldots, \mathbf{S}_{i+1}^{(t)}$), it follows that

$$\begin{aligned}&\mathbf{S}_{i+2}, \mathbf{S}_{[1,i+1]}, \mathbf{S}_{[1,i+1]}^{(1)}, \ldots, \mathbf{S}_{[1,i+1]}^{(t)}, \mathbf{R}_{[1,i+1]}, \mathbf{R}_{[1,i+1]}^{(1)}, \ldots, \mathbf{R}_{[1,i+1]}^{(t)}, \mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)} \\ \approx_{\epsilon_{i+1}+2\epsilon} &U_m, \mathbf{S}_{[1,i+1]}, \mathbf{S}_{[1,i+1]}^{(1)}, \ldots, \mathbf{S}_{[1,i+1]}^{(t)}, \mathbf{R}_{[1,i+1]}, \mathbf{R}_{[1,i+1]}^{(1)}, \ldots, \mathbf{R}_{[1,i+1]}^{(t)}, \mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)}.\end{aligned}$$

Also, we maintain at each step that $(\mathbf{X}, \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(t)})$ is independent from $(\mathbf{Y}, \mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(t)})$. This completes the proof. □

□

3.2 The Flip-Flop Primitive

The flip-flop primitive (Algorithm 1), introduced by Cohen [Coh15a], is a particularly elegant way of using the alternating extraction protocol. In this section, we establish a property of the flip-flip primitive that strengthens a result proved in [Coh15a]. Before presenting the flip-flop construction

and our result, we first discuss a toy example which motivates this construction.

A Toy Problem Let $\mathbf{X}, \mathbf{X}'$ be correlated r.v's such that $\mathbf{X}$ is an (n, k)-source. Further suppose we have access to a uniform strong $\mathbf{Y}$ on d bits such that $\{\mathbf{X}, \mathbf{X}'\}$ is independent of $\mathbf{Y}$. Our goal is to construct a deterministic function $f : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ such that

$$f(\mathbf{X}, \mathbf{Y}), f(\mathbf{X}', \mathbf{Y}) \approx \mathbf{U}_m, f(\mathbf{X}', \mathbf{Y}).$$

We adopt the following notation for convenience: For any r.v $\mathbf{Z} = g(\mathbf{X}, \mathbf{Y})$, where g is a deterministic function, let $\mathbf{Z}' = g(\mathbf{X}', \mathbf{Y})$. As a starting point for our construction of f, we could play an alternating extraction game, say for 2 rounds, between $\mathbf{X}$ and $\mathbf{Y}$ (and similarly, in the 'tampered game' between $\mathbf{X}'$ and $\mathbf{Y}'$). Let $\mathbf{R}_0, \mathbf{R}_1$ be the output of the look-ahead extractor. As a preliminary candidate for f, define $f(\mathbf{X}, \mathbf{Y}) = \mathbf{R}_1$. Thus, assuming k is large enough, we know by results from the previous section that $\mathbf{R}_1$ is close to uniform on average conditioned on $\mathbf{R}_0, \mathbf{R}'_0$. However, it is not clear if $\mathbf{R}_1$ is close to uniform given $\mathbf{R}'_1 = f(\mathbf{X}', \mathbf{Y})$. In fact it is not hard to find counter-examples for this construction. Thus, maybe we could try to solve an easier problem. Suppose we now we have access to an 'advice' bit b, and a tampered bit $b' \neq b$, and we are aiming to construct $f : \{0,1\}^n \times \{0,1\}^d \times \{0,1\} \to \{0,1\}^m$ such that

$$f(\mathbf{X}, \mathbf{Y}, b), f(\mathbf{X}', \mathbf{Y}, b') \approx \mathbf{U}_m, f(\mathbf{X}', \mathbf{Y}', b').$$

We could now try to define $f(\mathbf{X}, \mathbf{Y}, b) = \mathbf{R}_b$. Clearly, if $b = 1$ (and this $b' = 0$), this works since $\mathbf{R}_1$ is close to uniform on average given $\mathbf{R}'_0$. However, the construction does not work if $b = 0$. We give a rough idea of how to fix this approach and refer the reader to Algorithm 1 for the actual cosntruction. The idea is play 2 more rounds of alternating extraction between $\mathbf{X}$ and $\overline{\mathbf{Y}}$, where $\overline{\mathbf{Y}}$ is a new source derived from $\mathbf{Y}$ (and hence is independent of $\mathbf{X}$). Let $\overline{\mathbf{R}}_0, \overline{\mathbf{R}}_1$ be the output of the look-ahead extractor. We now define $f(\mathbf{X}, \mathbf{Y}) = \overline{\mathbf{R}}_{1-b}$. Thus, if $b = 0$, clearly $f(\mathbf{X}, \mathbf{Y}) = \overline{\mathbf{R}}_1$ is close to uniform on average given $f(\mathbf{X}', \mathbf{Y}) = \overline{\mathbf{R}}'_0$. Further, we can show that if $b = 1$, since we gain independence in the first 2 rounds of alternating extraction (i.e, $\mathbf{R}_1$ is close to uniform on average

given $\mathbf{R}'_0$), this carries on to the next two rounds of alternating extraction as well.

We now present the flip-flop construction, and then establish our result (Lemma 3.2.1) which informally states that the correlations are broken even allowing multiple tamperings on both $\mathbf{X}$ and $\mathbf{Y}$.

Algorithm 1: flip-flop(x, y, q_i, b)

Input: Bit strings x, y, q_i of length n_w, n_y, n_q respectively, and a bit b.
Output: A bit string of length n_q.
Subroutine: Let $\text{Ext}_q : \{0,1\}^{n_q} \times \{0,1\}^m \to \{0,1\}^m$ be a strong seeded extractor set to extract from min-entropy k with error ϵ and seed length m. Let $\text{Ext}_w : \{0,1\}^{n_w} \times \{0,1\}^m \to \{0,1\}^m$ be a strong seeded extractor set to extract from min-entropy k with error ϵ and seed length d.
Let $\text{laExt} : \{0,1\}^{n_w} \times \{0,1\}^{n_q+m} \to \{0,1\}^{2m}$ be the look ahead extractors defined in Section 3.1 for an alternating extraction protocol with parameters $m, u = 2$ (recall u is the number of steps in the protocol, m is the length of each random variable that is communicated between the players), and using $\text{Ext}_q, \text{Ext}_w$ as the strong seeded extractors.
Let $\text{Ext} : \{0,1\}^{n_y} \times \{0,1\}^m \to \{0,1\}^{n_q}$ be a strong seeded extractor set to extract from min-entropy k_1 with error ϵ.

1 Let $s_{i,1} = \text{Slice}(q_i, m)$
2 Let $\text{laExt}(x, (q_i, s_{i,1})) = r_{i,0}, r_{i,1}$
3 Let $\overline{q_i} = \text{Ext}(y, r_{i,b})$
4 Let $\overline{s_{i,1}} = \text{Slice}(\overline{q_i}, m)$.
5 Let $\text{laExt}(x, (\overline{q_i}, \overline{s_{i,1}})) = \overline{r_{i,1}}, \overline{r_{i,2}}$.
6 Let $q_{i+1} = \text{Ext}(y, \overline{r_{i,1-b}})$
7 Ouput q_{i+1}.

Lemma 3.2.1. *Let $b, \{b^h : h \in [j]\}$ be $j+1$ bits such that for all $h \in [j]$, $b \neq b^h$. Let $\mathbf{X}$ be a (n_w, k_w)-source and let $\{\mathbf{X}^h : h \in [j]\}$ be random variables on $\{0,1\}^{n_w}$ that are arbitrarily correlated with $\mathbf{X}$. Let $\mathbf{Y}, \{\mathbf{Y}^h : h \in [j]\}$ be arbitrarily correlated random variables that are independent of $(\mathbf{X}, \{\mathbf{X}^h : h \in [j]\})$. Suppose that $\mathbf{Y}$ is a (n_y, k_y)-source, $k_y = n_y - \lambda$, each random variable in $\{\mathbf{Y}^h : h \in [j]\}$ is on n_y bits. Let $\mathbf{Q}_i$ be some function of $\mathbf{Y}$ on n_q bits with min-entropy at least $n_q - \lambda$, and for each $h \in [j]$, let $\mathbf{Q}^h$ be an an arbitrary function of $\mathbf{Y}, \{\mathbf{Y}^a : a \in [j]\}$ on n_q bits.*

Let flip-flop *be the function computed by Algorithm* 1. *Let* flip-flop$(\mathbf{X}, \mathbf{Y}, \mathbf{Q}_i, b) = \mathbf{Q}_{i+1}$, *and for $h \in [j]$, let* flip-flop$(\mathbf{X}^h, \mathbf{Y}^h, \mathbf{Q}^h_i, b^h) = \mathbf{Q}^h_{i+1}$. *Suppose $k_y \geq \max\{k, k_1\} + 10\left(jn_q + jm + \log\left(\frac{1}{\epsilon}\right)\right)$,*

$k_w \geq k + 10\left(jm + \log\left(\frac{1}{\epsilon}\right)\right)$, *and* $n_q \geq k + 10jm + 2\log(\frac{1}{\epsilon}) + \lambda$.

Then, with probability at least $1 - \epsilon'$, *where* $\epsilon' = O(2^\lambda \epsilon)$, *over the fixing of the random variables* $\mathbf{Q}_i, \{\mathbf{Q}_i^h : h \in [j]\}, \mathbf{R}_{i,0}, \mathbf{R}_{i,1}, \{\mathbf{R}_{i,0}^h, \mathbf{R}_{i,1}^h : h \in [j]\}, \overline{\mathbf{Q}}_i, \{\overline{\mathbf{Q}}_i^h : h \in [j]\}, \overline{\mathbf{R}}_{i,0}, \overline{\mathbf{R}}_{i,1}, \{\overline{\mathbf{R}}_{i,0}^h, \overline{\mathbf{R}}_{i,1}^h : h \in [j]\}, \{\mathbf{Q}_{i+1}^h : h \in [j]\}$*:*

- $\mathbf{Q}_{i+1}$ *is* ϵ'*-close to* U_{n_q} *and is a deterministic function of* $\mathbf{Y}$
- *The random variables* $(\mathbf{X}, \{\mathbf{X}^h : h \in [j]\})$ *and* $(\mathbf{Y}, \{\mathbf{Y}^h : h \in [j]\})$ *are independent*
- $\mathbf{X}$ *has min-entropy at least* $k_w - 10\left(jm + \log\left(\frac{1}{\epsilon}\right)\right)$ *and* $\mathbf{Y}$ *has min-entropy at least* $k_y - 10\left(jn_q + jm + \log\left(\frac{1}{\epsilon}\right)\right)$.

Proof. **Notation:** For any determinitic function f, if $\mathbf{V} = f(\mathbf{X}, \mathbf{Y})$, let $\mathbf{V}^a$ denote the random variable $H(\mathbf{X}^a, \mathbf{Y}^a)$.

We split the proof into two cases, depending on b.

Case 1: Suppose $b = 1$. By Lemma 3.1.1, it follows that

$$\begin{aligned} &\mathbf{R}_{i,1}, \{\mathbf{R}_{i,0}^h : h \in [j]\}, \mathbf{Q}_i, \{\mathbf{Q}_i^h : h \in [j]\} \\ \approx_{\epsilon_1} &U_m, \{\mathbf{R}_{i,0}^h : h \in [j]\}, \mathbf{Q}_i, \{\mathbf{Q}_i^h : h \in [j]\}, \end{aligned}$$

where $\epsilon_1 = c2^\lambda\epsilon$, for some constant c. Thus, we can fix $\{\mathbf{R}_{i,0}^h : h \in [j]\}, \mathbf{Q}_i, \{\mathbf{Q}_i^h : h \in [j]\}$, and with probability at least $1 - O(\epsilon_1)$, $\mathbf{R}_{i,1}$ is $O(\epsilon_1)$-close to U_m. Note that $\mathbf{R}_{i,1}$ is now a deterministic function of $\mathbf{X}$. Further, by Lemma 2.3.7, $\mathbf{Y}$ loses min-entropy at most $(j+1)n_q + \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1 - \epsilon$ due to this fixing. Since on fixing $\mathbf{Q}_i, \{\mathbf{Q}_i^h : h \in [j]\}$, the random variables $\{\mathbf{R}_{i,0}^h : h \in [j]\}$ are deterministic function of $\mathbf{X}, \{\mathbf{X}^h : h \in [j]\}$, the source $\mathbf{X}$ loses min-entropy at most $jm + \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1 - \epsilon$ due to this fixing. We now note that the random variables $\{\overline{\mathbf{Q}}_i^h : h \in [j]\}$ are deterministic functions of $\mathbf{Y}, \{\mathbf{Y}^h : h \in [j]\}$. Thus, we fix $\{\overline{\mathbf{Q}}_i^h : h \in [j]\}$, and by Lemma 2.3.7, $\mathbf{Y}$ loses min-entropy at most $jn_q + \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1 - \epsilon$ due to this fixing. Since Ext extracts from min-entropy k_1, and k_y was chosen large enough, it follows that the random variable $\overline{\mathbf{Q}}_i$ is $(\epsilon + \epsilon_1)$-close to U_{n_q} with probability at least $1 - O(\epsilon_1)$ even after

the fixing. Further, we fix $\mathbf{R}_{i,1}$ since Ext is a strong seeded extractor, and by Lemma 2.3.7, $\mathbf{X}$ loses min-entropy at most $m + \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1-\epsilon$ due to this fixing. Thus $\overline{\mathbf{Q}}_i$ is now a deterministic function of $\mathbf{Y}$. We now fix the random variables $\{\mathbf{R}^h_{i,2} : h \in [j]\}$, noting that they are deterministic functions of $\mathbf{X}$ and hence does not affect the distribution of $\overline{\mathbf{Q}}_i$. $\mathbf{X}$ loses min-entropy at most $jm + \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1-\epsilon$ due to this fixing.

We now note that the random variables $\{\overline{\mathbf{R}}^h_{i,0}, \overline{\mathbf{R}}^h_{i,1} : h \in [j]\}$ are deterministic function of $\mathbf{X}, \{\mathbf{X}^j : j \in [h]\}$ since we have fixed $\{\overline{\mathbf{Q}}^{(h)}_i : h \in [j]\}$. Thus, we can fix $\{\overline{\mathbf{R}}^{(h)}_{i,0}, \overline{\mathbf{R}}^{(h)}_{i,1} : h \in [j]\}$ and $\mathbf{X}$ loses min-entropy at most $2jm + \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1-\epsilon$. Thus it follows by Lemma 3.1.1 that $|\overline{\mathbf{R}}_{i,1}, \overline{\mathbf{Q}}_i - U_m, \overline{\mathbf{Q}}_i| < \epsilon + O(\epsilon_1)$. We fix $\overline{\mathbf{Q}}_i$ and $\mathbf{Y}$ loses min-entropy at most $n_q + \log\left(\frac{1}{\epsilon}\right)$ using Lemma 2.3.7. Finally, we note that $\{\mathbf{Q}^h_{i+1} : h \in [j]\}$ is now a deterministic function of $\mathbf{Y}, \{\mathbf{Y}^h : h \in [j]\}$. Thus, we can fix $\{\mathbf{Q}^h_{i+1} : h \in [j]\}$ variables and $\mathbf{Y}$ loses min-entropy at most $jn_q + \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1-\epsilon$ due to this fixing. Further, $\overline{\mathbf{R}}_{i,0}$ is now a deterministic function of $\mathbf{X}$. It follows that $\mathbf{Q}_{i+1}$ is $O(\epsilon_1 + \epsilon)$-close to U_{n_q} since k_y is chosen large enough. We further fix $\overline{\mathbf{R}}_{i,1}$ noting that Ext is a strong extractor and $\mathbf{X}$ loses min-entropy at most $m + \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1-\epsilon$ due to this fixing.

Case 2: Now suppose $b = 0$. We fix the random variables $\mathbf{Q}_i, \{\mathbf{Q}^h_i : h \in [j]\}$. Conditioned on this fixing, it follows by Lemma 3.1.1 that $|\mathbf{R}_{i,0} - U_m| < \epsilon_1$, $\epsilon_1 = O(2^\lambda \epsilon)$, with probability at least $1-\epsilon$. Since Ext is a strong seeded extractor (and k_y is large enough) and $\mathbf{R}_{i,0}$ is a deterministic function of $\mathbf{X}$, it follows that $|\overline{\mathbf{Q}}_i, \mathbf{R}_{i,0} - U_{n_q}, \mathbf{R}_{i,0}| < \epsilon + \epsilon_1$ with probability at least ϵ. We fix $\mathbf{R}_{i,0}$, and observe that $\overline{\mathbf{Q}}_i$ is now a deterministic function of $\mathbf{Y}$. We can now fix $\{\mathbf{R}^h_{i,0}, \mathbf{R}^h_{i,1} : h \in [j]\}$ since $\{\mathbf{R}^h_{i,1} : h \in [j]\}$ is a deterministic function of $\mathbf{X}, \{\mathbf{X}^h : h \in [j]\}$, and hence does not affect the distribution of $\overline{\mathbf{Q}}_i$. As a result of these fixings, it is clear that $(\mathbf{X}, \{\mathbf{X}^h : h \in [j]\})$ is independent of $(\mathbf{Y}_i, \{\mathbf{Y}^h : h \in [j]\})$. Further $\mathbf{X}$ loses min-entropy of at most $2(j+1)m + \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1-\epsilon$, and $\mathbf{Y}$ loses min-entropy of at most $2(j+1)n_q + (j+1)m + 3\log\left(\frac{1}{\epsilon}\right)$ with probability at least $1-3\epsilon$. Note that now $\overline{\mathbf{Q}}_i, \{\mathbf{Q}^h_i : h \in [j]\}$ are deterministic functions of $\mathbf{Y}, \{\mathbf{Y}^h : h \in [j]\}$,

and $\overline{\mathbf{Q}}_i$ is $O(\epsilon_1)$-close to U_{n_q}. By Lemma 3.1.1, it follows that

$$\overline{\mathbf{R}}_{i,1}, \{\overline{\mathbf{R}}^h_{i,0} : h \in [j]\}, \overline{\mathbf{Q}}_i, \{\overline{\mathbf{Q}}^h_i : h \in [j]\} \approx_{\epsilon_2} U_m, \{\overline{\mathbf{R}}^h_{i,0} : h \in [j]\}, \overline{\mathbf{Q}}_i, \{\overline{\mathbf{Q}}^h_i : h \in [j]\}$$

where $\epsilon_2 = c(\epsilon_1 + \epsilon + \epsilon)$, for some constant c. Thus, we can fix $\{\overline{\mathbf{R}}^h_{i,0} : h \in [j]\}, \overline{\mathbf{Q}}_i, \{\overline{\mathbf{Q}}^h_i : h \in [j]\}$ and with probability at least $1 - O(\epsilon_2)$, $\overline{\mathbf{R}}_{i,1}$ is $O(\epsilon_2)$-close to U_m. Note that $\overline{\mathbf{R}}_{i,1}$ is now a deterministic function of $\mathbf{X}$. Further, by Lemma 2.3.7, $\mathbf{Y}$ loses min-entropy at most $(j+1)n_q + \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1 - \epsilon$ due to this fixing. Since on fixing $\overline{\mathbf{Q}}_i, \{\overline{\mathbf{Q}}^h_i : h \in [j]\}$, the random variables $\{\overline{\mathbf{R}}^h_{i,1} : h \in [j]\}$ are deterministic functions of $\mathbf{X}, \{\mathbf{X}^{(h)} : h \in [j]\}$, the source $\mathbf{X}$ loses min-entropy at most $jm + \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1 - \epsilon$ due to this fixing. We now note that the random variables $\{\mathbf{Q}^h_{i+1} : h \in [j]\}$ are deterministic functions of $\mathbf{Y}, \{\mathbf{Y}^h : h \in [j]\}$. Thus, we fix $\{\mathbf{Q}^h_{i+1} : h \in [j]\}$ and by Lemma 2.3.7, $\mathbf{Y}$ loses min-entropy at most $(j+1)n_q + \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1 - \epsilon$ due to this fixing. Since Ext extracts from min-entropy k_1, (and k_y is large enough) it follows that random variable $\mathbf{Q}_{i+1}$ is $O(\epsilon_2)$-close to U_{n_q} even after the fixing. Further, we fix $\overline{\mathbf{R}}_{i,1}$ since Ext is a strong seeded extractor, and by Lemma 2.3.7, $\mathbf{X}$ loses min-entropy $m + \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1 - \epsilon$ due to this fixing. Further $\mathbf{Q}_{i+1}$ is now a deterministic function of $\mathbf{Y}$. Thus we can fix the random variables $\{\overline{\mathbf{R}}^{(h)}_{i,2} : h \in [j]\}$ since they are deterministic function sod $\mathbf{X}$ and does not affect the distribution of $\mathbf{Q}_{i+1}$. $\mathbf{X}$ loses min-entropy at most $m + \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1 - \epsilon$ due to this fixing. This completes the proof. □

3.3 Correlation Breakers with Advice

In this section, we construct a primitive that breaks correlations under a weaker guarantee compared to the flip-flop function. Informally, as motivated in the previous section, the bit b can be thought of as advice to the flip-flop function, with the guarantee that $b \neq b^i$ for any i. Instead, now suppose we only have access to a short string w, with the guarantee that $w \neq w^i$. We show that it is possible to break correlations with this weaker guarantee by chaining together a bunch of flip-flop functions. This generalizes an object introduced by Cohen [Coh15a], which he called a local

correlation breaker, with our twist being that we now allow access to an advice string w. We now describe the construction in more detail.

Algorithm 2: ACB(x, y, z)
Input: Bit strings x, y, z of length n_w, n_y, ℓ respectively.
Output: A bit string of length n_q.

1 Let $q_1 = \text{Slice}(y, n_q)$
2 **for** $h = 1$ *to* ℓ **do**
3 $\quad q_{h+1} = 2\text{laExt}(x, y, q_h, z_h)$
4 **end**
5 Ouput $q_{\ell+1}$.

Lemma 3.3.1. *Let $z, z^1, \ldots, z^t$ each be ℓ bit strings such that for all $i \in [t]$, $z \neq z^i$. Let $\mathbf{X}$ be a (n_w, k_w)-source and let $\mathbf{X}^1, \ldots, \mathbf{X}^t$ be random variables on $\{0,1\}^{n_w}$ that are arbitrarily correlated with $\mathbf{X}$. Let $\mathbf{Y}, \mathbf{Y}^1, , \ldots, \mathbf{Y}^t$ be random variables on n_y bits that are independent of $(\mathbf{X}, \mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^t)$. Suppose that $\mathbf{Y}$ is a (n_y, k_y)-source, $k_y = n_y - \lambda$.*

Let ACB *be the function computed by Algorithm 2. Let* ACB$(\mathbf{X}, \mathbf{Y}, z) = \mathbf{Q}_{\ell+1}$, *and for $h \in [t]$, let* ACB$(\mathbf{X}^h, \mathbf{Y}^h, z^h) = \mathbf{Q}^h_{\ell+1}$. *Suppose $k_y \geq \max\{k, k_1\} + 20\ell\left(tn_q + tm + \log\left(\frac{1}{\epsilon}\right)\right)$, $k_w \geq k + 20\ell\left(tm + \log\left(\frac{1}{\epsilon}\right)\right)$ and $n_q \geq k + 10tm + 2\log(\frac{1}{\epsilon}) + \lambda$. Then, we have*

$$\mathbf{Q}_{\ell+1}, \mathbf{Q}^1_{\ell+1}, \ldots, \mathbf{Q}^t_{\ell+1} \approx_{\epsilon'} U_{n_q}, \mathbf{Q}^1_{\ell+1}, \ldots, \mathbf{Q}^t_{\ell+1}$$

where $\epsilon'_\ell = O((2^\lambda + \ell)\epsilon)$.

Proof. **Notation:** For any function f, if $V = f(\mathbf{X}, \mathbf{Y})$, let V^a denote the random variable $f(\mathbf{X}^{(a)}, \mathbf{Y}^{(a)})$.

For $h \in [\ell]$, define the sets

$$\text{Ind}_h = \{i \in [t] : z_h \neq z^i_h, \qquad \overline{\text{Ind}}_h = [t] \setminus \text{Ind}_h,$$

$$\text{Ind}_{[h]} = \cup_{i=1}^h \text{Ind}_h, \qquad \overline{\text{Ind}}_{[h]} = [t] \setminus \text{Ind}_{[h]}.$$

We record a simple claim.

Claim 3.3.2. *For each $i \in [t]$, there exists $h \in [\ell]$ such that $i \in \mathrm{Ind}_h$.*

Proof. Recall that we have fixed $\mathbf{Z}, \mathbf{Z}^1, \ldots, \mathbf{Z}^t$ such that $\mathbf{Z} \neq \mathbf{Z}^i$ for any $i \in [t]$. Thus it follows that for each $i \in [t]$, there exists some $h \in [\ell]$ such that $\mathbf{Z}_h \neq \mathbf{Z}^i_h$, and hence $i \in \mathrm{Ind}_h$. □

We now prove our main claim, which combined with Lemma 3.2.1 and a simple inductive argument proves Lemma 3.3.1.

Claim 3.3.3. *For any $h \in \{0, 1, \ldots, \ell\}$, suppose the following holds:*

With probability at least $1 - \epsilon_h$ over the fixing of the random variables $\{\mathbf{Q}_i : i \in [h]\}, \{\mathbf{Q}^j_i : i \in [h], j \in [t]\}, \{\mathbf{R}_{i,1}, \mathbf{R}_{i,2} : i \in [h]\}, \{\mathbf{R}^j_{i,1}, \mathbf{R}^j_{i,2} : i \in [h], j \in [t]\}, \{\overline{\mathbf{Q}}_i : i \in [h]\}, \{\overline{\mathbf{Q}}^j_i : i \in [h], j \in [t]\}, \{\overline{\mathbf{R}}_{i,0}, \overline{\mathbf{R}}_{i,1} : i \in [h]\}, \{\overline{\mathbf{R}}^j_{i,0}, \overline{\mathbf{R}}^{(j)}_{i,1} : i \in [h], j \in [t]\}, \{\mathbf{Q}^j_{i+1} : j \in \mathrm{Ind}_{[h]}\}$: (a) $\mathbf{Q}_{h+1}$ is ϵ_h-close to a source with min-entropy at least $n_q - \lambda$ and is a deterministic function of $\mathbf{Y}$ (b) $\{\mathbf{Q}^j_{h+1} : j \in \overline{\mathrm{Ind}}_{[h]}\}$ is a deterministic function of $\mathbf{Y}, \{\mathbf{Y}^j : j \in [t]\}$ (c) The random variables $(\mathbf{X}, \{\mathbf{X}^j : j \in [t]\})$ and $(\mathbf{Y}, \{\mathbf{Y}^j : j \in [t]\})$ are independent (d) $\mathbf{X}$ has min-entropy at least $k_w - 10h\left(tm + \log\left(\frac{1}{\epsilon}\right)\right) > k + 10\left(tm + \log\left(\frac{1}{\epsilon}\right)\right)$ and $\mathbf{Y}$ has min-entropy at least $k_y - 10h\left(tn_q + tm + \log\left(\frac{1}{\epsilon}\right)\right) > \max\{k, k_1\} + 10\left(tn_q + tm + \log\left(\frac{1}{\epsilon}\right)\right)$.

Then, the following holds:

Let $\epsilon_{h+1} = \epsilon_h + c2^\lambda\epsilon$ for some constant c. With probability at least $1 - \epsilon_{h+1}$ over the fixing of the random variables $\{\mathbf{Q}_i : i \in [h+1]\}, \{\mathbf{Q}^j_i : i \in [h+1], j \in [t]\}, \{\mathbf{R}_{i,1}, \mathbf{R}_{i,2} : i \in [h+1]\}, \{\mathbf{R}^j_{i,1}, \mathbf{R}^j_{i,2} : i \in [h+1], j \in [t]\}, \{\overline{\mathbf{Q}}_i : i \in [h+1]\}, \{\overline{\mathbf{Q}}^j_i : i \in [h+1], j \in [t]\}, \{\overline{\mathbf{R}}_{i,0}, \overline{\mathbf{R}}_{i,1} : i \in [h]\}, \{\overline{\mathbf{R}}^j_{i,0}, \overline{\mathbf{R}}^j_{i,1} : i \in [h+1], j \in [t]\}, \{\mathbf{Q}^j_{i+1} : j \in \mathrm{Ind}_{[h+1]}\}$: (a) $\mathbf{Q}_{h+2}$ is ϵ_{h+1}-close to $\mathbf{U}_{n_q}$ and is a deterministic function of $\mathbf{Y}$ (b) $\{\mathbf{Q}^j_{h+2} : j \in \overline{\mathrm{Ind}}_{[h+1]}\}$ is a deterministic function of $\mathbf{Y}, \{\mathbf{Y}^j : j \in [t]\}$ (c) The random variables $(\mathbf{X}, \{\mathbf{X}^j : j \in [t]\})$ and $(\mathbf{Y}, \{\mathbf{Y}^j : j \in [t]\})$ are independent (d) $\mathbf{X}$ has min-entropy at least $k_w - 10(h+1)\left(tm + \log\left(\frac{1}{\epsilon}\right)\right)$ and $\mathbf{Y}$ has min-entropy at least $k_y - 10(h+1)\left(tm + \log\left(\frac{1}{\epsilon}\right)\right)$.

Proof. We fix the random variables $\{\mathbf{Q}_i : i \in [h]\}, \{\mathbf{Q}^j_i : i \in [h], j \in [t]\}, \{\mathbf{R}_{i,0}, \mathbf{R}_{i,1} : i \in [h]\}, \{\mathbf{R}^j_{i,0}, \mathbf{R}^j_{i,1} : i \in [h], j \in [t]\}, \{\overline{\mathbf{Q}}_i : i \in [h]\}, \{\overline{\mathbf{Q}}^j_i : i \in [h], j \in [t]\}, \{\overline{\mathbf{R}}_{i,0}, \overline{\mathbf{R}}_{i,1} : i \in [h]\}, \{\overline{\mathbf{R}}^j_{i,0}, \overline{\mathbf{R}}^j_{i,1} :$

$i \in [h], j \in [t]\}, \{\mathbf{Q}_{i+1}^j : j \in \mathrm{Ind}_{[h]}\}$ such that (a), (b), (c), (d) holds (this happens with probability at least $1-\epsilon_h$. We also fix the random variables $\{\mathbf{R}^j_{h+1,\psi_1(z^j_{h+1})} : j \in \mathrm{Ind}_{[h]}\}$, noting that they are deterministic functions of $\mathbf{X}$. Thus $\mathbf{X}$ has min-entropy at least $k_w - 10h\left(jm + \log\left(\frac{1}{\epsilon}\right)\right) - tm - \log\left(\frac{1}{\epsilon}\right)$ with probability at least $1-\epsilon$. Further, $\mathbf{Q}$ has min-entropy at least $k_y - 10h\left(tn_q + tm + \log\left(\frac{1}{\epsilon}\right)\right)$. The claim now follows directly from Lemma 3.2.1. □

To complete the proof of Lemma 3.3.1, we now note that the hypothesis of Claim 3.3.3 is indeed satisfied when $h = 0$. Thus, by ℓ applications of Claim 3.3.3, it follows that the $\mathbf{Q}_{\ell+1}$ is ϵ'_ℓ-close to U_{n_q}, where $\epsilon'_\ell = O(2^\lambda\epsilon + \ell\epsilon)$. This follows since for all applications of Claim 3.3.3 except the first time, $\mathbf{Q}_h$ is ϵ_h-close to uniform, and hence the parameter $\lambda = 0$. This concludes the proof of Lemma 3.3.1. □

3.4 Handling Linear Correlations

In the above sections, we crucially use the fact that $\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t$ is independent of $\mathbf{Y}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t$. In this section, we show that in fact this can be relaxed and we can handle some amount of 'linear correlation' among these r.v's. We now describe the setting in more details. Let $\mathbf{Y}^1, \ldots, \mathbf{Y}^t$ be correlated random variables. We show that it is possible to break the correlations by just using an additional correlated source of the form $\mathbf{X} + \mathbf{Z}$, assuming $\mathbf{X}$ is independent of $\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t$ (and $\mathbf{Z}$ is allowed to have arbitrary correlations with $\mathbf{Y}^1, \ldots, \mathbf{Y}^t$).

The main idea is to adapt the methods from the previous section with an important change. We now use linear seeded extractors in the alternating extraction steps to exploit the linearity of the correlations between the source and the seed in various steps of the protocol. The proofs of the results in this section are similar to that of the Section 3.2. However, to carry out the arguments requires more careful conditioning and a slightly subtler inductive hypothesis in some of the proofs.

We begin by proving a result similar to Lemma 3.1.1 when an alternating extraction protocol is run between the sources $\mathbf{W} = \mathbf{X} + \mathbf{Z}$ and $\mathbf{Q} = \mathbf{Y}$, where $\mathbf{Y}$ and $\mathbf{Z}$ are arbitrarily correlated and $\mathbf{X}$ is independent of $(\mathbf{Y}, \mathbf{Z})$.

Lemma 3.4.1. *For any $\epsilon > 0$ and any integers $n_1, n_2, k, k_1, t, d, h$ satisfying $k_1 \geq k + 2(t+1)d(h+1) + \log(1/\epsilon)$ and $n_2 \geq k + 2(t+1)d(h+1) + \log(1/\epsilon)$, let*

- $\mathbf{X}$ *be an (n_1, k_1)-source, $\mathbf{Y} = \mathbf{U}_{n_2}$ and $\mathbf{Z}$ be a random variable on n_1 bits.*
- $\mathbf{Y}^1, \ldots, \mathbf{Y}^t$ *be random variables on n_2 bits each, such that $\mathbf{X}$ is independent of $\{\mathbf{Y}, \mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.*
- $\mathbf{S}_0 = \text{Slice}(\mathbf{Y}, d)$ *and for $i \in [t]$, $\mathbf{S}_0^i = \text{Slice}(\mathbf{Y}^i, d)$.*
- $\text{LExt}_1 : \{0,1\}^{n_1} \times \{0,1\}^d \to \{0,1\}^d$ *and* $\text{LExt}_2 : \{0,1\}^{n_2} \times \{0,1\}^d \to \{0,1\}^d$ *be (k, ϵ)-strong linear seeded extractors.*
- $\text{laExt}(\mathbf{X} + \mathbf{Z}, (\mathbf{Y}, \mathbf{S}_0)) = \mathbf{R}_1, \ldots, \mathbf{R}_h$, *and for $i \in [t]$,* $\text{laExt}(\mathbf{X} + \mathbf{Z}, (\mathbf{Y}^i, \mathbf{S}_0^i)) = \mathbf{R}_1^i, \ldots, \mathbf{R}_h^i$, *where* laExt *is executed with the linear seeded extractors* $\text{LExt}_1, \text{LExt}_2$ *for h rounds.*
- $\mathbf{R}_{j,\mathbf{X}} = \text{LExt}_1(\mathbf{X}, \mathbf{S}_j)$ *and* $\mathbf{R}_{j,\mathbf{Z}} = \text{LExt}_1(\mathbf{Z}, \mathbf{S}_j)$, *$j \in [0, h]$.*

Then,

1. *for any $j \geq 0$,*

$$\begin{aligned} &\mathbf{S}_j, \{\mathbf{S}_g : g \in [0, j-1]\}, \{\mathbf{S}_g^i : g \in [0, j-1], i \in [t]\}, \{\mathbf{R}_g : g \in [0, j-1]\}, \\ &\qquad \{\mathbf{R}_g^i : g \in [0, j-1], i \in [t]\} \\ \approx_{(4j+2)\epsilon}\ &\mathbf{U}_d, \{\mathbf{S}_g : g \in [0, j-1]\}, \{\mathbf{S}_g^i : g \in [0, j-1], i \in [t]\}, \{\mathbf{R}_g : g \in [0, j-1]\}, \\ &\qquad \{\mathbf{R}_g^i : g \in [0, j-1], i \in [t]\}. \end{aligned}$$

2. *for any $j \geq 0$, conditioned on $\{\mathbf{S}_g : g \in [0, j-1]\}, \{\mathbf{S}_g^i : g \in [0, j-1], i \in [t]\}, \{\mathbf{R}_g : g \in [0, j-1]\}, \{\mathbf{R}_g^i : g \in [0, j-1], i \in [t]\}$,*
   - $\mathbf{X}$ *is independent of $\{\mathbf{Y}, \mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.*
   - $\mathbf{S}_j$ *and $\{\mathbf{S}_j^i : i \in [t]\}$ are deterministic functions of $\mathbf{Y}$.*

- $\mathbf{X}$ *has conditional min-entropy at least* $k + (t+1)d(h+1-j) + \log(1/\epsilon)$ *and* $\mathbf{Y}$ *has conditional min-entropy at least* $k + 2(t+1)d(h+1-j) + \log(1/\epsilon)$.

3. *for any* $j \geq 0$,

$$\begin{aligned} &\mathbf{R}_j, \mathbf{R}_{j,\mathbf{Z}}, \{\mathbf{R}_g : g \in [0, j-1]\}, \{\mathbf{R}^i_{j,\mathbf{Z}} : i \in [t]\}, \{\mathbf{R}^i_g : g \in [0, j-1], i \in [t]\}, \\ &\qquad \{\mathbf{S}_g : g \in [0, j]\}, \{\mathbf{S}^i_g : g \in [0, j], i \in [t]\}, \mathbf{Y}, \{\mathbf{Y}^i : i \in [t]\}, \mathbf{Z} \\ \approx_{4(j+1)\epsilon}\ &\mathbf{U}_d, \mathbf{R}_{j,\mathbf{Z}}, \{\mathbf{R}_g : g \in [0, j-1]\}, \{\mathbf{R}^i_{j,\mathbf{Z}} : i \in [t]\}, \{\mathbf{R}^i_g : g \in [0, j-1], i \in [t]\}, \\ &\qquad \{\mathbf{S}_g : g \in [0, j]\}, \{\mathbf{S}^i_g : g \in [0, j], i \in [t]\}, \mathbf{Y}, \{\mathbf{Y}^i : i \in [t]\}, \mathbf{Z}. \end{aligned}$$

4. *for any* $j \geq 0$, *conditioned on* $\mathbf{R}_{j,\mathbf{Z}}, \{\mathbf{R}_g : g \in [0, j-1]\}, \{\mathbf{R}^i_{j,\mathbf{Z}} : i \in [t]\}, \{\mathbf{R}^i_g : g \in [0, j-1], i \in [t]\}, \{\mathbf{S}_g : g \in [0, j]\}, \{\mathbf{S}^i_g : g \in [0, j], i \in [t]\}$,

   - $\mathbf{X}$ *is independent of* $\{\mathbf{Y}, \mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.
   - $\mathbf{R}_j$ *and* $\{\mathbf{R}^i_j : i \in [t]\}$ *are deterministic function of* $\mathbf{X}$.
   - $\mathbf{X}$ *has conditional min-entropy at least* $k + (t+1)d(h+1-j) + \log(1/\epsilon)$ *and* $\mathbf{Y}$ *has conditional min-entropy at least* $k + 2(t+1)d(h-j) + \log(1/\epsilon)$.

Proof. We prove the lemma by induction on j. The validity of the lemma when $j = 0$ is direct. Thus, suppose that the lemma holds for $j-1$ for some $j \in [h]$ and we prove it for j.

Fix the following random variables:

$$\begin{aligned} &\mathbf{R}_{j-1,\mathbf{Z}}, \{\mathbf{R}_g : g \in [0, j-2]\}, \{\mathbf{R}^i_{j-1,\mathbf{Z}} : i \in [t]\}, \{\mathbf{R}^i_g : g \in [0, j-2], i \in [t]\}, \\ &\qquad \{\mathbf{S}_g : g \in [0, j-1]\}, \{\mathbf{S}^i_g : g \in [0, j-1], i \in [t]\}. \end{aligned}$$

By induction hypothesis, it follows that

- $\mathbf{R}_{j-1}$ is $4j\epsilon$-close to $\mathbf{U}_d$ on average and is a deterministic function of $\mathbf{X}$.
- $\mathbf{Y}$ has conditional min-entropy $k + 2(t+1)d(h+1-j) + \log(1/\epsilon)$ and is independent of $\mathbf{X}$.

- $\mathbf{X}$ has conditional min-entropy $k + (t+1)d(h+2-j) + \log(1/\epsilon)$.

since $\mathbf{S}_j = \mathrm{LExt}_2(\mathbf{Y}, \mathbf{R}_{j-1})$, it follows by Lemma 2.2.5 that $\mathbf{S}_j$ is $(4j+2)\epsilon$-close to $\mathbf{U}_d$ on average conditioned on $\mathbf{R}_{j-1}$. Thus we fix $\mathbf{R}_{j-1}$ and observe that $\mathbf{S}_j$ is now a deterministic function of $\mathbf{Y}$. Next we fix $\{\mathbf{R}^i_{j-1} : i \in [t]\}$ observing that, by induction hypothesis, they are deterministic functions of $\mathbf{X}$ and hence does not affect $\mathbf{S}_j$. As a result of this fixing, $\{\mathbf{S}^i_j : i \in [t]\}$ is now a deterministic function of $\mathbf{Y}$, and further $\mathbf{X}$ remains independent of $\{\mathbf{Y}, \mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. We note that all the random variables fixed in this step are deterministic functions of $\mathbf{X}$. Thus after these fixings, by Lemma 2.2.3 and induction hypothesis, the conditional entropy of $\mathbf{X}$ is at least $k + (t+1)d(h+2-j) - (t+1)d + \log(1/\epsilon) = k + (t+1)d(h+1-j) + \log(1/\epsilon)$. This concludes the proof of (1) and (2).

We now prove (3) and (4). We continue to condition on the random variables that we have fixed so far in our proof. We have,

- $\mathbf{S}_j$ is $(4j+2)\epsilon$-close to $\mathbf{U}_d$ on average and is a deterministic function of $\mathbf{Y}$,
- $\mathbf{X}$ has average conditional min-entropy at least $k + (t+1)(h+1-j) + \log(1/\epsilon)$ and is independent of $\mathbf{Y}$,
- $\mathbf{Y}$ has conditional min-entropy $k + 2(t+1)d(h+1-j) + \log(1/\epsilon)$.

Thus, it follows by Lemma 2.2.5 that $\mathbf{R}_{j,\mathbf{X}} = \mathrm{LExt}_1(\mathbf{X}, \mathbf{S}_j)$ is $4(j+1)\epsilon$-close to $\mathbf{U}_d$ on average conditioned on $\mathbf{S}_j$. We fix $\mathbf{S}_j$ and note that $\mathbf{R}_{j,\mathbf{X}}$ is now a deterministic function of $\mathbf{X}$. Next, we fix $\mathbf{R}_{j,\mathbf{Z}}$ which is now a deterministic function of $\mathbf{Z}$ and hence does not affect $\mathbf{R}_{j,\mathbf{X}}$. Since LExt_1 is linear seeded, it follows that $\mathbf{R}_j = \mathbf{R}_{j,\mathbf{X}} + \mathbf{R}_{j,\mathbf{Z}}$ and $\mathbf{R}^i_j = \mathbf{R}^i_{j,\mathbf{X}} + \mathbf{R}^i_{j,\mathbf{Z}}$. Thus $\mathbf{R}_j$ is ϵ_j-close to $\mathbf{U}_d$ on average and is a deterministic function of $\mathbf{X}$. We now fix $\{\mathbf{S}^i_j : i \in [t]\}$ which is a deterministic function of $\mathbf{Y}$, and next fix $\{\mathbf{R}^i_{j,\mathbf{Z}} : i \in [t]\}$ which is a deterministic function of $\mathbf{Z}$. Thus, these additional fixings do not affect $\mathbf{R}_j$. Finally observe that $\mathbf{X}$ remains independent of $\{\mathbf{Y}, \mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. We note that all the random variables fixed in this step are deterministic functions of $\{\mathbf{Y}, \mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. Thus after these fixings, by Lemma 2.2.3, the conditional entropy

of $\mathbf{Y}$ is at least $k + 2(t+1)d(h+1-j) - 2(t+1)d + \log(1/\epsilon) = k + 2(t+1)d(h-j) + \log(1/\epsilon)$. This concludes the proof of induction and hence the lemma follows. □

We now instantiate the flip-flop and Advice-Correlation Breaker functions with linear seeded extractors.

Algorithm 3: flip-flop(y^i, y^i_j, w, b)

Input: Bit strings $y^i, y^i_j, w = x + z$ of length n_1, n_2, n_1 respectively, and a bit b.
Output: Bit string y^i_{j+1} of length n_2.
Subroutines: Let $\text{LExt}_1 : \{0,1\}^{n_1} \times \{0,1\}^d \to \{0,1\}^d$, $\text{LExt}_2 : \{0,1\}^{n_2} \times \{0,1\}^d \to \{0,1\}^d$ be (k, ϵ)-strong linear seeded extractors. Let $\text{LExt}_3 : \{0,1\}^{n_1} \times \{0,1\}^d \to \{0,1\}^{n_2}$ be a (k_2, ϵ)-strong linear seeded extractor.
Let $\text{laExt} : \{0,1\}^{n_1} \times \{0,1\}^{n_2+d} \to \{0,1\}^{2d}$ be a look-ahead extractor for an alternating extraction protocol run for 2 rounds using $\text{LExt}_1, \text{LExt}_2$ as the seeded extractors.

1 Let $s^i_{0,j} = \text{Slice}(y^i_j, d)$, $\text{laExt}(w, (y^i_j, s^i_{0,j})) = r^i_{0,j}, r^i_{1,j}$
2 Let $\overline{y^i_{1,j}} = \text{LExt}_3(y^i, r^i_{b,j})$
3 Let $\overline{s^i_{0,j}} = \text{Slice}(\overline{y^i_{1,j}}, d)$, $\text{laExt}(w, (\overline{y^i_{1,j}}, \overline{s^i_{0,j}})) = \overline{r^i_{0,j}}, \overline{r^i_{1,j}}$
4 Output $y^i_{j+1} = \text{LExt}_3(y^i, \overline{r^i_{1-b,j}})$

Algorithm 4: ACB(y^i, w, id)

Input: Bit strings $y^i, w = x + z, id$ of length n_1, n_1, h respectively.
Output: Bit string y_{h+1} of length n_2.

1 Let $y^i_1 = \text{Slice}(y, n_2)$
2 **for** $j = 1$ *to* h **do**
3 $\quad y^i_{j+1} = \text{flip-flop}(y^i, y^i_j, w, id[j])$
4 **end**
5 Output y^i_{h+1}.

Theorem 3.4.2. *For any $\epsilon > 0$ and any integers $n_1, n_2, k, k_1, t, d, h$ satisfying $k_1 \geq k + 8tdh + \log(1/\epsilon)$, $n_2 \geq k + 3td + \log(1/\epsilon)$, $n_1 \geq k + 10tdh + (4ht+1)n_2 + \log(1/\epsilon)$, let*

- $\mathbf{X}$ *be an (n_1, k_1)-source, $\mathbf{Y}^1 = \mathbf{U}_{n_1}$ and $\mathbf{Z}, \mathbf{Y}^2, \ldots, \mathbf{Y}^t$ be random variables on n_1 bits each, such that $\mathbf{X}$ is independent of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.*

- $id^1, \ldots, id^t$ *be bit strings of length* h *such that for each* $i \in [t]$, $id^1 \neq id^i$.
- $\mathbf{Y}^i_{h+1} = \mathrm{ACB}(\mathbf{Y}, \mathbf{X} + \mathbf{Z}, id^i)$ *for* $i \in [t]$ *where* ACB *is the function computed by Algorithm* 4.

Then,

$$\mathbf{Y}^1_{h+1}, \mathbf{Y}^2_{h+1}, \ldots, \mathbf{Y}^t_{h+1} \approx_{O(h\epsilon)} \mathbf{U}_{n_2}, \mathbf{Y}^2_{h+1}, \ldots, \mathbf{Y}^t_{h+1}.$$

Proof. Define the following sets for $j \in [h]$:

$$\mathrm{Ind}_j = \{i \in [2,h] : id^i[j] \neq id^1[j]\}, \quad \mathrm{Ind}_{\leq j} = \cup_{g=1}^{j} \mathrm{Ind}_g, \quad \overline{\mathrm{Ind}_{\leq j}} = [t] \setminus \mathrm{Ind}_{\leq j}.$$

We prove the following lemma from which Theorem 3.4.2 is direct by observing that $\mathrm{Ind}_{\leq h} = [2,t]$.

Lemma 3.4.3. *For each* $j \in [h]$,

$$\mathbf{Y}^1_{j+1}, \{\mathbf{Y}^i_{j+1} : i \in \mathrm{Ind}_{\leq j}\} \approx_{O(j\epsilon)} \mathbf{U}_{n_2}, \{\mathbf{Y}^i_{j+1} : i \in \mathrm{Ind}_{\leq j}\}.$$

Proof. Recall that $\mathbf{R}_{c,j} = \mathrm{LExt}(\mathbf{X} + \mathbf{Z}, \mathbf{S}_{c,j})$ (for any $c \in \{0,1\}$ and $j \in [h]$). Define $\mathbf{R}_{c,j,\mathbf{X}} = \mathrm{LExt}(\mathbf{X}, \mathbf{S}_{c,j})$ and $\mathbf{R}_{c,j,\mathbf{Z}} = \mathrm{LExt}(\mathbf{Z}, \mathbf{S}_{c,j})$. Since LExt is linear seeded, it follows that $\mathbf{R}_{c,j} = \mathbf{R}_{c,j,\mathbf{X}} + \mathbf{R}_{c,j,\mathbf{Z}}$. Similarly, define $\overline{\mathbf{R}_{c,j,\mathbf{X}}} = \mathrm{LExt}(\mathbf{X}, \overline{\mathbf{S}_{c,j}})$ and $\overline{\mathbf{R}_{c,j,\mathbf{Z}}} = \mathrm{LExt}(\mathbf{Z}, \overline{\mathbf{S}_{c,j}})$.

We prove the lemma by induction on j. In fact, we prove the following stronger statement:

For every $j \in [0,h]$, conditioned on the random variables: $\{\mathbf{Y}^i_{j+1} : i \in \mathrm{Ind}_{\leq j}\}, \{\mathbf{Y}^i_g : g \in [j], i \in [t]\}, \{\mathbf{R}^i_{0,j+1,\mathbf{Z}} : i \in \mathrm{Ind}_j\}, \{\overline{\mathbf{Y}}^i_g : g \in [j], i \in [t]\}, \{\mathbf{S}^i_{0,g} : g \in [j], i \in [t]\}, \{\mathbf{S}^i_{1,g} : g \in [j], i \in [t]\}, \{\mathbf{R}^i_{0,g} : g \in [j], i \in [t]\}, \{\mathbf{R}^i_{1,g} : g \in [j], i \in [t]\}, \{\overline{\mathbf{S}^i_{0,g}} : g \in [j], i \in [t]\}, \{\overline{\mathbf{S}^i_{1,g}} : g \in [j], i \in [t]\}, \{\overline{\mathbf{R}^i_{0,g}} : g \in [j], i \in [t]\}, \{\overline{\mathbf{R}^i_{1,g}} : g \in [j], i \in [t]\}$

- $\mathbf{Y}^1_{j+1}$ is $6j\epsilon$-close to $\mathbf{U}_{n_2}$ on average
- $\mathbf{X}$ is independent of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.
- $\{\mathbf{Y}^i_{j+1} : i \in [t]\}$ is a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.

- $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} = k+8td(h-j)+\log(1/\epsilon)$ and $\mathbf{Y}^1$ has conditional min-entropy at least $k_{j,\mathbf{Y}} = k + 10td(h-j) + 4tn_2(h-j+1) + \log(1/\epsilon)$.

The base case of the induction when $j = 0$ is direct. Now suppose the above holds for some $j-1 \geq 0$, and we prove it for j.

We fix the following random variables: $\{\mathbf{Y}^i_j : i \in \mathrm{Ind}_{\leq(j-1)}\}, \{\mathbf{Y}^i_g : g \in [j-1], i \in [t]\}, \{\overline{\mathbf{Y}}^i_g : g \in [j-1], i \in [t]\}, \{\mathbf{R}^i_{0,j,\mathbf{Z}} : i \in \mathrm{Ind}_{j-1}\}, \{\mathbf{S}^i_{0,g} : g \in [j-1], i \in [t]\}, \{\mathbf{S}^i_{1,g} : g \in [j-1], i \in [t]\}, \{\mathbf{R}^i_{0,g} : g \in [j-1], i \in [t]\}, \{\mathbf{R}^i_{1,g} : g \in [j-1], i \in [t]\}, \{\overline{\mathbf{S}^i_{0,g}} : g \in [j-1], i \in [t]\}, \{\overline{\mathbf{S}^i_{1,g}} : g \in [j-1], i \in [t]\}, \{\overline{\mathbf{R}^i_{0,g}} : g \in [j-1], i \in [t]\}, \{\overline{\mathbf{R}^i_{1,g}} : g \in [j-1], i \in [t]\}$. By induction hypothesis, we have

- $\mathbf{Y}^1_j$ is $6(j-1)\epsilon$-close to $\mathbf{U}_{n_2}$ on average.
- $\mathbf{X}$ is independent of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.
- $\{\mathbf{Y}^i_j : i \in [t]\}$ is a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.
- $\mathbf{X}$ has conditional min-entropy at least $k_{j-1,\mathbf{X}} = k_{j,\mathbf{X}} + 8td$ and $\mathbf{Y}^1$ has conditional min-entropy at least $k_{j-1,\mathbf{Y}} = k_{j,\mathbf{Y}} + 10td + 4tn_2$.

We repeatedly use Lemma 2.2.5 when we argue about the remaining conditional min-entropy in a random variable and do not explicitly mention this. Further, any random variable that we fix is either a deterministic function of $\mathbf{X}$ or a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. Thus, we always maintain that $\mathbf{X}$ is independent of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$ and again do not explicitly mention this.

We split the proof into two cases depending on the bit $id^1[j]$.

Case 1: Suppose $id^1[j] = 1$ and hence $\overline{\mathbf{Y}}^1_j = \mathrm{LExt}_3(\mathbf{Y}^1, \mathbf{R}^1_{1,j})$. It follows that for all $i \in \mathrm{Ind}_j$, $id^i[j] = 0$ and $\overline{\mathbf{Y}}^i_j = \mathrm{LExt}_3(\mathbf{Y}^i, \mathbf{R}^i_{0,j})$. Since $\{\mathbf{Y}^i_j : i \in \mathrm{Ind}_{\leq(j-1)}\}$ is fixed, it follows that for all $i \in \mathrm{Ind}_{\leq(j-1)}$, $\mathbf{R}^i_{0,j,\mathbf{X}} = \mathrm{LExt}_1(\mathbf{X}, \mathbf{S}^i_{0,j})$ is a deterministic function of $\mathbf{X}$. We fix the random variables $\{\mathbf{R}^i_{0,j,\mathbf{X}} : i \in \mathrm{Ind}_{\leq(j-1)}\}$, and $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} + 7td$. We now fix $\mathbf{S}^1_{0,j}$, $\{\mathbf{S}^i_{0,j} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$, $\{\mathbf{R}^i_{0,j,\mathbf{Z}} : i \in [t]\}$ and by Lemma 3.4.1, it follows that (a) $\mathbf{R}^1_{0,j}$ is $(6j-5)\epsilon$-close to uniform on average and is a deterministic function of $\mathbf{X}$, (b) $\mathbf{X}$ has conditional

min-entropy at least $k_{j,\mathbf{X}} + 7td$ and $\mathbf{Y}^1_j$ has conditional min-entropy at least $k + td + \log(1/\epsilon)$. We also note that for each $i \in \mathrm{Ind}_{\leq(j-1)}$, $\mathbf{R}^i_{0,j} = \mathbf{R}^i_{0,j,\mathbf{X}} + \mathbf{R}^i_{0,j,\mathbf{Z}}$ is fixed.

Next we fix $\{\mathbf{S}^i_{1,j} : i \in \mathrm{Ind}_{\leq(j-1)}\}$, observing that it is now a deterministic function of $\{\mathbf{Y}^i : i \in [t]\}$ and hence does not affect the distribution of $\mathbf{R}^1_{0,j}$. The conditional min-entropy of $\mathbf{Y}^1_j$ after this fixing is at least $k + \log(1/\epsilon)$. We now fix $\mathbf{R}^1_{0,j}$, $\{\mathbf{R}^i_{0,j} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$ and by Lemma 3.4.1, (a) $\mathbf{S}^1_{1,j}$ is $(6j-4)\epsilon$-close to uniform on average and is a deterministic function of $\mathbf{Y}^1$, (b) $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} + 6td$ and $\mathbf{Y}^1_j$ has conditional min-entropy at least $k + \log(1/\epsilon)$.

Continuing in a similar fashion as above, we first fix $\{\mathbf{R}^i_{1,j,\mathbf{X}} : i \in \mathrm{Ind}_{\leq(j-1)}\}$, which is a deterministic function of $\mathbf{X}$. The conditional min-entropy of $\mathbf{X}$ after this fixing is at least $k_{j,\mathbf{X}} + 5td$. We now fix the random variables $\mathbf{S}^1_{1,j}, \{\mathbf{S}^i_{1,j} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$, $\{\mathbf{R}^i_{1,j,\mathbf{Z}} : i \in [t]\}$, $\{\mathbf{Y}^i_j : i \in [t]\}$ and by Lemma 3.4.1, we have (a) $\mathbf{R}^1_{1,j}$ is $(6j-3)\epsilon$-close to uniform on average and is a deterministic function of $\mathbf{X}$, (b) $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} + 5td$.

We fix $\{\overline{\mathbf{Y}}^i_j : i \in \mathrm{Ind}_{\leq(j-1)}\}$ which is deterministic function of $\{\mathbf{Y}^i : i \in [t]\}$, and $\mathbf{R}^1_{1,j}$ continues to remain close to $\mathbf{U}_d$ on average. We also fix $\{\overline{\mathbf{Y}}^i_j : i \in \mathrm{Ind}_j\}$ observing that it is a deterministic function of $\{\mathbf{Y}^i : i \in [t]\}$ (since we have fixed $\{\mathbf{R}^i_{0,j} : i \in [t]\}$ and for $i \in \mathrm{Ind}_j$, $\mathbf{Y}^i_j = \mathrm{LExt}_3(\mathbf{Y}^i, \mathbf{R}^i_{0,j})$). It follows that $\{\overline{\mathbf{S}}^i_{0,j} : i \in \mathrm{Ind}_{\leq j}\}$ is fixed and hence $\{\overline{\mathbf{R}}^i_{0,j,\mathbf{Z}} : i \in \mathrm{Ind}_{\leq j}\}$ is a deterministic function of $\mathbf{Z}$. Thus, we fix $\{\overline{\mathbf{R}}^i_{0,j,\mathbf{Z}} : i \in \mathrm{Ind}_{\leq j}\}$ without affecting the distribution of $\mathbf{R}^1_{1,j}$.

The conditional min-entropy of $\mathbf{Y}^1$ after this fixing is at least $k_{j,\mathbf{Y}} + 2tn_2 + 4td$. Thus $\overline{\mathbf{Y}}^1_j = \mathrm{LExt}_3(\mathbf{Y}^1, \mathbf{R}^1_{1,j})$ is $(6j-2)\epsilon$-close to $\mathbf{U}_{n_2}$ on average conditioned on $\mathbf{R}^1_{1,j}$. We fix $\mathbf{R}^1_{1,j}$ and thus $\overline{\mathbf{Y}}^1_j$ is now a deterministic function of $\mathbf{Y}^1$. We now fix $\{\mathbf{R}^i_{1,j,\mathbf{X}} : i \in \overline{\mathrm{Ind}_j}\}$ which is a deterministic function of $\mathbf{X}$ and note that this fixes $\{\mathbf{R}^i_{1,j} : j \in \overline{\mathrm{Ind}_j}\}$. Further, since $\{\overline{\mathbf{Y}}^i_j : i \in \mathrm{Ind}_{\leq j}\}$ is fixed, it follows that for all $i \in \mathrm{Ind}_{\leq j}$, $\overline{\mathbf{R}}^i_{0,j,\mathbf{X}}$ is a deterministic function of $\mathbf{X}$. We fix the random variables $\{\overline{\mathbf{R}}^i_{0,j,\mathbf{X}} : i \in \mathrm{Ind}_{\leq j}\}$ and note that $\{\overline{\mathbf{S}}^i_{1,j} : i \in \mathrm{Ind}_{\leq j}\}$ is now fixed. Thus $\{\overline{\mathbf{R}}^i_{1,j,\mathbf{X}} : i \in \mathrm{Ind}_{\leq j}\}$ is now a deterministic of $\mathbf{X}$. We fix $\{\overline{\mathbf{R}}^i_{1,j,\mathbf{X}} : i \in \mathrm{Ind}_{\leq j}\}$ and $\overline{\mathbf{Y}}^1_j$ continues to remain close to uniform on average and $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} + 2td$.

We now fix $\overline{\mathbf{S}}^1_{0,j}$, $\{\overline{\mathbf{S}}^i_{0,j} : i \in \overline{\text{Ind}_{\leq j}}\}$, $\{\overline{\mathbf{R}}^i_{0,j,\mathbf{Z}} : i \in \overline{\text{Ind}_{\leq j}}\}$, $\{\overline{\mathbf{Y}}^i_j : i \in \overline{\text{Ind}_{\leq j}}\}$ and by Lemma 3.4.1, it follows that (a) $\overline{\mathbf{R}}^1_{0,j}$ is $(6j-1)\epsilon$-close to uniform on average and is a deterministic function of $\mathbf{X}$, (b) $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} + 3td$. Next we fix $\{\overline{\mathbf{R}}^i_{1,j,\mathbf{Z}} : i \in \text{Ind}_{\leq j}\}$ which a deterministic function of $\mathbf{Z}$ and $\{\mathbf{Y}^i_{j+1} : i \in \text{Ind}_{\leq j}\}$ is now a deterministic function of $\{\mathbf{Y}^i : i \in \text{Ind}_{\leq j}\}$. Thus, we fix $\{\mathbf{Y}^i_{j+1} : i \in \text{Ind}_{\leq j}\}$ and $\overline{\mathbf{R}}^1_{0,j}$ continues to remain uniform on average. It now follows that $\{\mathbf{R}^i_{0,j+1,\mathbf{Z}} : i \in \text{Ind}_{\leq (j)}\}$ is a deterministic function of $\mathbf{Z}$, and we fix it.

The conditional min-entropy of $\mathbf{Y}^1$ after this fixing is at least $k_{j,\mathbf{Y}}$ and thus, $\mathbf{Y}^1_{j+1} = \text{LExt}_3(\mathbf{Y}^1, \overline{\mathbf{R}}^1_{0,j})$ is $6j\epsilon$-close to $\mathbf{U}_{n_2}$ on average conditioned on $\overline{\mathbf{R}}^1_{0,j}$. We fix $\overline{\mathbf{R}}^1_{0,j}$ which is a deterministic function of $\mathbf{X}$ and thus $\mathbf{Y}^1_{j+1}$ is now a deterministic function of $\mathbf{Y}^1$. Now consider any $i \in \overline{\text{Ind}_{\leq j}}$. since we have fixed $\overline{\mathbf{R}}^i_{0,j,\mathbf{Z}}$ and $\overline{\mathbf{R}}^i_{0,j} = \overline{\mathbf{R}}^i_{0,j,\mathbf{X}} + \overline{\mathbf{R}}^i_{0,j,\mathbf{Z}}$, it follows that $\overline{\mathbf{R}}^i_{0,j,\mathbf{X}}$ is a deterministic function of $\mathbf{X}$. Thus, we fix $\{\overline{\mathbf{R}}^i_{0,j} : i \in \overline{\text{Ind}_{\leq j}}\}$ without affecting the distribution of $\mathbf{Y}^1_{j+1}$. $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}} + td$ after this fixing. Now, since $\overline{\mathbf{Y}}^i_j$ is fixed, it follows that $\overline{\mathbf{S}}^1_{1,j}$ is fixed for each $i \in [t]$. Thus, for any $i \in \overline{\text{Ind}_{\leq j}}$, $\overline{\mathbf{R}}^i_{1,j,\mathbf{X}} = \text{LExt}_1(\mathbf{X}, \overline{\mathbf{S}}^i_{1,j})$ is a deterministic function of $\mathbf{X}$. We fix $\{\overline{\mathbf{R}}^i_{1,j,\mathbf{X}} : i \in \overline{\text{Ind}_{\leq j}}\}$, and observe that $\mathbf{Y}^1_{j+1}$ remains close to uniform on average and $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}}$. Thus, $\{\overline{\mathbf{R}}^i_{1,j} : i \in \overline{\text{Ind}_{\leq j}}\}$ is now a deterministic function of $\mathbf{Z}$ and $\{\mathbf{Y}^i_{j+1} : i \in \overline{\text{Ind}_{\leq j}}\}$ is a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. This concludes the proof of this case.

Case 2: Suppose $id^1[j] = 0$ and hence $\overline{\mathbf{Y}}^1_j = \text{LExt}_3(\mathbf{Y}^1, \mathbf{R}^1_{0,j})$. Since $\{\mathbf{Y}^i_j : j \in \text{Ind}_{j-1}\}$ is fixed, it follows that $\{\mathbf{R}^i_{0,j,\mathbf{X}} : i \in \text{Ind}_{\leq (j-1)}\}$ and $\{\mathbf{R}^i_{1,j,\mathbf{X}} : i \in \text{Ind}_{\leq (j-1)}\}$ are deterministic functions of $\mathbf{X}$ and we fix them without affecting the distribution of $\mathbf{Y}^1_j$. $\mathbf{X}$ has conditional min-entropy at least $k_{j-1,\mathbf{X}} + 6td$ after this fixing.

We now fix $\mathbf{S}^1_{0,j}, \{\mathbf{S}^i_{0,j} : i \in \overline{\text{Ind}_{\leq (j-1)}}\}$, $\mathbf{R}^1_{0,j,\mathbf{Z}}$, $\{\mathbf{R}^i_{0,j,\mathbf{Z}} : i \in \overline{\text{Ind}_{\leq (j-1)}}\}$ and by Lemma 3.4.1, $\mathbf{R}^1_{0,j}$ is $(6j-5)\epsilon$-close to $\mathbf{U}_d$ on average and is a deterministic function of $\mathbf{X}$. We next fix $\{\mathbf{R}^i_{1,j,\mathbf{Z}} : i \in \text{Ind}_{\leq (j-1)}\}$, $\{\mathbf{Y}^i_j : i \in \overline{\text{Ind}_{\leq (j-1)}}\}$, and $\{\overline{\mathbf{Y}}^i_j : i \in \text{Ind}_{\leq (j-1)}\}$ observing that they are deterministic functions of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$ and does not affect the distribution of $\mathbf{R}^1_{0,j}$. Further, $\{\overline{\mathbf{R}}^i_{0,j,\mathbf{Z}} : i \in \text{Ind}_{\leq (j-1)}\}$ is now a deterministic function of $\mathbf{Z}$, and we fix it.

As a result of these fixings, $\mathbf{Y}^1$ has conditional min-entropy at least $k_{j-1,\mathbf{Y}} + 5tdh + 2tn_2$.

Thus, $\overline{\mathbf{Y}}^1_j$ is $(6j-4)\epsilon$-close to $\mathbf{U}_{n_2}$ on average conditioned on $\mathbf{R}^1_{0,j}$. We fix $\mathbf{R}^1_{0,j}$ and $\overline{\mathbf{Y}}^1_j$ is now a deterministic function of $\mathbf{Y}^1$. We now fix $\{\mathbf{R}^i_{0,j,\mathbf{X}} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$ which is a deterministic function of $\mathbf{X}$ and note that this fixes $\{\mathbf{S}^i_{0,j} : i \in \mathrm{Ind}_{\leq(j-1)}\}$. Thus $\{\mathbf{R}^i_{1,j,\mathbf{X}} : i \in \mathrm{Ind}_{\leq(j-1)}\}$ is now a deterministic function of $\mathbf{X}$ and we fix it without affecting the distribution of $\overline{\mathbf{Y}}^i_j$. As a result of this fixing $\{\mathbf{R}^i_{1,j} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$ is a deterministic function of $\mathbf{Z}$ and hence $\{\overline{\mathbf{Y}}^i_j : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$ is a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. Next, we fix $\{\overline{\mathbf{R}}^i_{0,j} : i \in \mathrm{Ind}_{\leq(j-1)}\}$ and $\{\overline{\mathbf{R}}^i_{1,j} : i \in \mathrm{Ind}_{\leq(j-1)}\}$, noting that they are deterministic functions of $\mathbf{X}$. $\mathbf{X}$ has conditional min-entropy at least $k_{j-1,\mathbf{X}} + 2td$ after these fixings.

We now fix $\overline{\mathbf{S}}^1_{0,j}$, $\{\overline{\mathbf{S}}^i_{0,j} : i \in \overline{\mathrm{Ind}_{\leq(h-1)}}\}$, $\overline{\mathbf{R}}^1_{0,j,\mathbf{Z}}$, $\{\overline{\mathbf{R}}^i_{0,j,\mathbf{Z}} i \in \overline{\mathrm{Ind}_{\leq(h-1)}}\}$ and invoking Lemma 3.4.1, it follows that $\overline{\mathbf{R}}^1_{0,j}$ is $(6j-3)\epsilon$-close to uniform on average and is a deterministic function of $\mathbf{X}$. We now fix $\{\overline{\mathbf{R}}^i_{1,j,\mathbf{Z}} : i \in \mathrm{Ind}_{\leq(j-1)}\}$ which a deterministic function of $\mathbf{Z}$ and note that this fixes $\{\overline{\mathbf{R}}^i_{1,j} : i \in \mathrm{Ind}_{\leq(j-1)}\}$. Further $\overline{\mathbf{Y}}^1_j$ has conditional min-entropy at least $k + td + \log(1/\epsilon)$. We now fix $\{\overline{\mathbf{R}}^i_{0,j,\mathbf{X}} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$, $\{\overline{\mathbf{S}}^i_{1,j,\mathbf{X}} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$, $\{\overline{\mathbf{R}}^i_{1,j,\mathbf{X}} : i \in \overline{\mathrm{Ind}_{\leq(j-1)}}\}$, and by Lemma 3.4.1, it follows that $\overline{\mathbf{R}}^i_{1,j}$ is $(6j-1)\epsilon$-close to $\mathbf{U}_d$ on average and is deterministic function of $\mathbf{X}$.

We now observe that $\{\mathbf{Y}^i_{j+1} : i \in \mathrm{Ind}_{\leq j}\}$ is a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$ and fix it without affecting the distribution of $\overline{\mathbf{R}}^1_{1,j}$. Next we fix $\{\mathbf{R}^i_{0,j,\mathbf{Z}} : i \in \mathrm{Ind}_{\leq j}\}$ which is now a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. The conditional min-entropy of $\mathbf{Y}^1$ is at least $k_{j,\mathbf{Y}}$ and hence $\mathbf{Y}^i_{j+1}$ is $6j\epsilon$-close to $\mathbf{U}_{n_2}$ on average conditioned on $\overline{\mathbf{R}}^1_{1,j}$. We fix $\overline{\mathbf{R}}^1_{1,j}$ and thus $\mathbf{Y}^1_{j+1}$ is now a deterministic function of $\mathbf{Y}^1$. Thus we fix $\{\overline{\mathbf{R}}^i_{1,j,\mathbf{X}} : i \in \overline{\mathrm{Ind}_{\leq j}}\}$ and as a result $\{\mathbf{Y}^i_{j+1} : i \in \overline{\mathrm{Ind}_{\leq j}}\}$ is now a deterministic function of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$. Further $\mathbf{X}$ has conditional min-entropy at least $k_{j,\mathbf{X}}$ as a result of these fixings. This completes the proof of induction and the theorem follows. □

□

3.5 Non-Malleable Independence Preserving Mergers

In this section, we construct a primitive to break correlations in an even more general setting. To motivate the general problem, consider the following simpler setting: Let $\mathbf{X}$ be a $2 \times n$ matrix r.v

with rows $\mathbf{X}_1$ and $\mathbf{X}_2$, and let $\mathbf{X}'$ be a correlated $2\times n$ matrix with rows $\mathbf{X}'_1$ and $\mathbf{X}'_2$. Further, suppose we know that either (a) $\mathbf{X}_1, \mathbf{X}'_1 \approx \mathbf{U}_n, \mathbf{X}'_1$ or (b) $\mathbf{X}_2, \mathbf{X}'_2 \approx \mathbf{U}_n, \mathbf{X}'_1$ holds. Our goal is to break the correlations between these matrices using access to an independent seed $\mathbf{Y}$ (the seed is tampered as well to $\mathbf{Y}'$). More specifically, we want to construct a function $f : \{0,1\}^{2n} \times \{0,1\}^d \to \{0,1\}^m$ such that

$$f(\mathbf{X}, \mathbf{Y}), f(\mathbf{X}', \mathbf{Y}') \approx \mathbf{U}_m, f(\mathbf{X}', \mathbf{Y}').$$

Informally, we call a function that satisfies the above guarantee to be a non-malleable independence preserving merger (NIPM). More formally, we define an NIPM in the following way.

Definition 3.5.1. *A $(L, t, d', \epsilon, \epsilon')$-NIPM $: \{0,1\}^{Lm} \times \{0,1\}^d \to \{0,1\}^{m_1}$ satisfies the following property. Suppose*

- *$\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t$ are r.v's, each supported on boolean $L \times m$ matrices s.t for any $i \in [L]$, $|\mathbf{X}_i - \mathbf{U}_m| \leq \epsilon$,*
- *$\{\mathbf{Y}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$ is independent of $\{\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t\}$, s.t $\mathbf{Y}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t$ are each supported on $\{0,1\}^d$ and $H_\infty(\mathbf{Y}) \geq d - d'$,*
- *there exists an $h \in [L]$ such that $|(\mathbf{X}_h, \mathbf{X}^1_h, \ldots, \mathbf{X}^t_h) - (\mathbf{U}_m, \mathbf{X}^1_h, \ldots, \mathbf{X}^t_h)| \leq \epsilon$,*

then

$$\begin{aligned}|(L,t,d',\epsilon,\epsilon')\text{-NIPM}((\mathbf{X},\mathbf{Y}),(L,t,d',\epsilon,\epsilon')\text{-NIPM}(\mathbf{X}^1,\mathbf{Y}^1),\ldots,(L,t,d',\epsilon,\epsilon')\text{-NIPM}(\mathbf{X}^t,\mathbf{Y}^t)\\ -\mathbf{U}_{m_1},(L,t,d',\epsilon,\epsilon')\text{-NIPM}(\mathbf{X}^1,\mathbf{Y}^1),\ldots,(L,t,d',\epsilon,\epsilon')\text{-NIPM}(\mathbf{X}^t,\mathbf{Y}^t)| \leq \epsilon'.\end{aligned}$$

Using our NIPM, we construct a standard IPM introduced in the work of Cohen and Schulman [CS16].

Definition 3.5.2. *A $(L, C, k, t, \epsilon, \epsilon')$-IPM $: \{0,1\}^{Lm} \times \{0,1\}^n \to \{0,1\}^{m_1}$ satisfies the following property. Suppose*

- $\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t$ *are r.v's, each supported on boolean* $L \times m$ *matrices s.t for any* $i \in [L]$, $|\mathbf{X}_i - \mathbf{U}_m| \leq \epsilon$,
- $\mathbf{Y}^1, \ldots, \mathbf{Y}^C$ *is an* (n, k)*-source, independent of* $\{\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t\}$.
- *there exists an* $h \in [L]$ *such that* $|(\mathbf{X}_h, \mathbf{X}^1_h, \ldots, \mathbf{X}^t_h) - (\mathbf{U}_m, \mathbf{X}^1_h, \ldots, \mathbf{X}^t_h)| \leq \epsilon$,

then

$$|(L, C, k, t, \epsilon, \epsilon')\text{-IPM}(\mathbf{X}, \mathbf{Y}), (L, C, k, t, \epsilon, \epsilon')\text{-IPM}(\mathbf{X}^1, \mathbf{Y}), \ldots, (L, C, k, t, \epsilon, \epsilon')\text{-NIPM}(\mathbf{X}^t, \mathbf{Y})$$
$$-\mathbf{U}_{m_1}, (L, C, k, t, \epsilon, \epsilon')\text{-IPM}(\mathbf{X}^1, \mathbf{Y}), \ldots, (L, C, k, t, \epsilon, \epsilon')\text{-IPM}(\mathbf{X}^t, \mathbf{Y})| \leq \epsilon'$$

The key differences between an NIPM and IPM are the following: The function IPM is allowed to have access to multiple independent sources instead of a seed $\mathbf{Y}$ to break the correlation between $\mathbf{X}$ and $\mathbf{X}'$. Further, these independent sources are not subject to any tampering.

3.5.1 ℓ-Non-Malleable Independence Preserving Merger

The following result presents our basic construction of an NIPM. The construction is based on extending the technique of alternating extraction in a new (but simple) way. We refer the reader to Chapter 4 for improved constructions of NIPM which uses the basic NIPM from this section in a black-box way. Further using these explicit NIPM constructions, we also give improved constructions of IPM (see Chapter 7).

Theorem 3.5.3. *There exist constants* $c_{3.5.3}, c'_{3.5.3} > 0$ *such that for all integers* $m, d, k_1, \ell > 0$ *and any* $\epsilon > 0$, *with* $m \geq d \geq k_1 > c_{3.5.3}\ell \log(n/\epsilon)$, *there exists an explicit function* ℓ-NIPM : $(\{0,1\}^m)^\ell \times \{0,1\}^d \to \{0,1\}^{m_1}$, $m_1 = 0.9(m - c_{3.5.3}\ell \log(m/\epsilon))$, *such that if the following conditions hold:*

- $\mathbf{X}_1, \ldots, \mathbf{X}_\ell$ *are r.v's s.t for all* $i \in [\ell]$, $|\mathbf{X}_i - \mathbf{U}_m| \leq \epsilon_1$, *and* $\mathbf{X}'_1, \ldots, \mathbf{X}'_\ell$ *are r.v's with each* $\mathbf{X}'_i$ *supported on* $\{0,1\}^m$.

- $\{\mathbf{Y}, \mathbf{Y}'\}$ *is independent of* $\{\mathbf{X}_1, \ldots, \mathbf{X}_t, \mathbf{X}'_1, \ldots, \mathbf{X}'_t\}$*, s.t the r.v's* $\mathbf{Y}, \mathbf{Y}'$ *are both supported on* $\{0,1\}^d$ *and* $H_\infty(\mathbf{Y}) \geq k_1$*.*
- *there exists an* $h \in [t]$ *such that* $|(\mathbf{X}_h, \mathbf{X}'_h) - (\mathbf{U}_m, \mathbf{X}'_h)| \leq \epsilon$*,*

then

$$|\ell\text{-NIPM}((\mathbf{X}_1, \ldots, \mathbf{X}_\ell), \mathbf{Y}), \ell\text{-NIPM}((\mathbf{X}'_1, \ldots, \mathbf{X}'_\ell), \mathbf{Y}'), \mathbf{Y}, \mathbf{Y}'$$
$$-\mathbf{U}_{m_1}, \ell\text{-NIPM}((\mathbf{X}'_1, \ldots, \mathbf{X}'_\ell), \mathbf{Y}'), \mathbf{Y}, \mathbf{Y}'| \leq c'_{3.5.3}\ell\epsilon$$

Our construction of NIPM is based on extending the method of alternating extraction in a new way.

ℓ-Alternating Extraction We extend the above technique by letting Quentin have access to ℓ sources $\mathbf{Q}_1, \ldots, \mathbf{Q}_\ell$ (instead of just $\mathbf{Q}$) and ℓ strong-seeded extractors $\{\text{Ext}_{q,i} : i \in [\ell]\}$ such that in the i'th round of the protocol, he uses $\mathbf{Q}_i$ to produce the r.v $\mathbf{S}_i = \text{Ext}_{q,i}(\mathbf{Q}_i, \mathbf{R}_i)$. More formally, the following sequence of r.v's is generated: $\mathbf{S}_1 = \text{Slice}(\mathbf{Q}_1, d), \mathbf{R}_1 = \text{Ext}_w(\mathbf{W}, \mathbf{S}_1), \mathbf{S}_2 = \text{Ext}_{q,2}(\mathbf{Q}_2, \mathbf{R}_1), \ldots, \mathbf{R}_{\ell-1} = \text{Ext}_w(\mathbf{Q}_{\ell-1}, \mathbf{S}_{\ell-1}), \mathbf{S}_\ell = \text{Ext}_{q,\ell}(\mathbf{Q}_\ell, \mathbf{R}_\ell)$. Define the look-ahead extractor

$$\ell\text{-laExt}((\mathbf{Q}_1, \ldots, \mathbf{Q}_\ell), \mathbf{W}) = \mathbf{S}_\ell.$$

We are now ready to prove Theorem 3.5.3.

Proof of Theorem 3.5.3. We instantiate the ℓ-look-ahead extractor described above with the following strong seeded extractors: Let $\text{Ext}_1 : \{0,1\}^m \times \{0,1\}^{d_1} \to \{0,1\}^{d_1}$, $\text{Ext}_2 : \{0,1\}^d \times \{0,1\}^{d_1} \to \{0,1\}^{d_1}$ and $\text{Ext}_3 : \{0,1\}^m \times \{0,1\}^{d_1} \to \{0,1\}^{m_1}$ be explicit strong-seeded from Theorem 2.1.2 designed to extract from min-entropy $m/2, k_1/4, m - c_{3.5.3}\ell\log(m/\epsilon)$ respectively, each with error ϵ. Thus $d_1 = c_{2.1.2}\log(m/\epsilon)$.

We think of each $\mathbf{X}_i$ being uniform, and add back an error $\epsilon_1\ell$ in the end.

For each $i \in [\ell - 1]$, let $\text{Ext}_{q,i} = \text{Ext}_1$, $\text{Ext}_{q,\ell} = \text{Ext}_3$ and $\text{Ext}_w = \text{Ext}_2$.

Define

$$\mathrm{NIPM}((\mathbf{X}_1,\ldots,\mathbf{X}_\ell),\mathbf{Y}) = \mathrm{laExt}((\mathbf{X}_1,\ldots,\mathbf{X}_\ell),\mathbf{Y}).$$

For any random variable $\mathbf{V} = f((\mathbf{X}_1,\ldots,\mathbf{X}_\ell),\mathbf{Y})$ (where f is an arbitrary deterministic function), let $\mathbf{V}' = f((\mathbf{X}'_1,\ldots,\mathbf{X}'_\ell),\mathbf{Y}')$.

We first prove the following claim.

Claim 3.5.4. *For any $j \in [h-1]$, conditioned on the r.v's $\{\mathbf{S}_i : i \in [j-1]\}, \{\mathbf{S}'_i : i \in [j-1]\}, \{\mathbf{R}_i : i \in [j-1]\}, \{\mathbf{R}'_i : i \in [j-1]\}$ the following hold:*

- *$\mathbf{S}_j$ is $2(j-1)\epsilon$-close to $\mathbf{U}_{d_1}$,*
- *$\mathbf{S}_j, \mathbf{S}'_j$ are deterministic functions of $\{\mathbf{X}_j, \mathbf{X}'_j\}$,*
- *for each $i \in [t]$, $\mathbf{X}_i$ has average conditional min-entropy at least $m - 2(j-1)d_1 - \log(1/\epsilon)$,*
- *$\mathbf{Y}$ has average conditional min-entropy at least $k_1 - 2(j-1)d_1 - \log(1/\epsilon)$,*
- *$\{\mathbf{X}_1,\ldots,\mathbf{X}_\ell,\mathbf{X}'_1,\ldots,\mathbf{X}'_\ell\}$ is independent of $\{\mathbf{Y},\mathbf{Y}'\}$.*

Further, conditioned on the r.v's $\{\mathbf{S}_i : i \in [j]\}, \{\mathbf{S}'_i : i \in [j]\}, \{\mathbf{R}_i : i \in [j-1]\}, \{\mathbf{R}'_i : i \in [j-1]\}$ the following hold:

- *$\mathbf{R}_j$ is $(2j-1)\epsilon$-close to $\mathbf{U}_d$,*
- *$\mathbf{R}_j, \mathbf{R}'_j$ are deterministic functions of $\{\mathbf{Y},\mathbf{Y}'\}$,*
- *for any $i \in [\ell]$, $\mathbf{X}_i$ has average conditional min-entropy at least $m - 2jd_1 - \log(1/\epsilon)$,*
- *$\mathbf{Y}$ has average conditional min-entropy at least $k_1 - 2(j-1)d_1 - \log(1/\epsilon)$,*
- *$\{\mathbf{X}_1,\ldots,\mathbf{X}_\ell,\mathbf{X}'_1,\ldots,\mathbf{X}'_\ell\}$ is independent of $\{\mathbf{Y},\mathbf{Y}'\}$.*

Proof. We prove the above by induction on j. The base case when $j = 1$ is direct. Thus suppose $j > 1$. Fix the r.v's $\{\mathbf{S}_i : i \in [j-1]\}, \{\mathbf{S}'_i : i \in [j-1]\}, \{\mathbf{R}_i : i \in [j-2]\}, \{\mathbf{R}'_i : i \in [j-2]\}$. Using inductive hypothesis, it follows that

- $\mathbf{R}_{j-1}$ is $(2j-3)\epsilon$-close to $\mathbf{U}_d$,
- $\mathbf{R}_{j-1}, \mathbf{R}'_{j-1}$ are deterministic functions of $\{\mathbf{Y}, \mathbf{Y}'\}$,
- for any $i \in [t]$, $\mathbf{X}_i$ has average conditional min-entropy at least $m - 2(j-1)d_1 - \log(1/\epsilon)$,
- $\mathbf{Y}$ has average conditional min-entropy at least $k_1 - 2(j-2)d_1 - \log(1/\epsilon)$,
- $\{\mathbf{X}_1, \ldots, \mathbf{X}_\ell, \mathbf{X}'_1, \ldots, \mathbf{X}'_\ell\}$ is independent of $\{\mathbf{Y}, \mathbf{Y}'\}$.

Now since $\mathbf{S}_j = \mathrm{Ext}_1(\mathbf{X}_j, \mathbf{R}_{j-1})$, it follows that $\mathbf{S}_j$ is $2(j-1)\epsilon$-close to $\mathbf{U}_{d_1}$ on average conditioned on $\mathbf{R}_{j-1}$. We thus fix $\mathbf{R}_{j-1}$. Further, we also fix $\mathbf{R}'_{j-1}$ without affecting the distribution of $\mathbf{S}_j$. Thus $\mathbf{S}_j, \mathbf{S}'_j$ are now a deterministic function of $\mathbf{X}_j, \mathbf{X}'_j$. It follows that after these fixings, the average conditional min-entropy of $\mathbf{Y}$ is at least $k_1 - 2(j-2)d_1 - \log(1/\epsilon) - 2d_1 = k_1 - 2(j-1)d_1 - \log(1/\epsilon)$.

Next, we have $\mathbf{R}_j = \mathrm{Ext}_2(\mathbf{Y}, \mathbf{S}_j)$, and thus fixing $\mathbf{S}_j$, it follows that $\mathbf{R}_j$ is $(2j-1)\epsilon$-close to uniform on average. Further, since $\mathbf{R}_j$ is now a deterministic function of $\mathbf{Y}$, we fix $\mathbf{S}'_j$. As a result of these fixings, each $\mathbf{X}_i$ loses conditional min-entropy at most $2d_1$ on average. Since at each point, we either fix a r.v that is a deterministic function of either $\{\mathbf{X}_1, \ldots, \mathbf{X}_\ell, \mathbf{X}'_1, \ldots, \mathbf{X}_\ell\}$ or $\{\mathbf{Y}, \mathbf{Y}'\}$ it follows that $\{\mathbf{X}_1, \ldots, \mathbf{X}_\ell, \mathbf{X}'_1, \ldots, \mathbf{X}'_\ell\}$ remain independent of $\{\mathbf{Y}, \mathbf{Y}'\}$. This completes the inductive step, and hence the proof follows. □

We now proceed to prove the following claim.

Claim 3.5.5. *Conditioned on the r.v's* $\{\mathbf{S}_i : i \in [h-1]\}, \{\mathbf{S}'_i : i \in [h]\}, \{\mathbf{R}_i : i \in [h-1]\}, \{\mathbf{R}'_i : i \in [h]\}$ *the following hold:*

- $\mathbf{S}_h$ *is* $2(h-1)\epsilon$*-close to* $\mathbf{U}_d$*,*
- $\mathbf{S}_h$ *is a deterministic function of* $\mathbf{X}_h$*,*
- *for each* $i \in [t]$, $\mathbf{X}_i$ *has average conditional min-entropy at least* $m - 2hd_1 - \log(1/\epsilon)$*,*
- $\mathbf{Y}$ *has average conditional min-entropy at least* $k_1 - 2hd_1 - \log(1/\epsilon)$*,*

- $\{\mathbf{X}_1, \ldots, \mathbf{X}_\ell, \mathbf{X}'_1, \ldots, \mathbf{X}'_\ell\}$ *is independent of* $\{\mathbf{Y}, \mathbf{Y}'\}$.

Proof. We fix the r.v's $\{\mathbf{S}_i : i \in [h-1]\}, \{\mathbf{S}'_i : i \in [h-1]\}, \{\mathbf{R}_i : i \in [h-2]\}, \{\mathbf{R}'_i : i \in [h-2]\}$, and using Claim 3.5.4 the following hold:

- $\mathbf{R}_{h-1}$ is $(2h-3)\epsilon$-close to $\mathbf{U}_d$,
- $\mathbf{R}_{h-1}, \mathbf{R}'_{h-1}$ are deterministic functions of $\{\mathbf{Y}, \mathbf{Y}'\}$,
- for any $i \in [t]$, $\mathbf{X}_i$ has average conditional min-entropy at least $m - 2(h-1)d_1 - \log(1/\epsilon)$,
- $\mathbf{Y}$ has average conditional min-entropy at least $k_1 - 2(h-2)d_1 - \log(1/\epsilon)$,
- $\{\mathbf{X}_1, \ldots, \mathbf{X}_\ell, \mathbf{X}'_1, \ldots, \mathbf{X}'_\ell\}$ is independent of $\{\mathbf{Y}, \mathbf{Y}'\}$.

Next we claim that $\mathbf{X}_h$ has average conditional min-entropy at least $m - 2(h-1)d_1 - \log(1/\epsilon)$ even after fixing $\mathbf{X}'_h$. We know that before fixings any other r.v, we have $\mathbf{X}_h|\mathbf{X}'_h$ is ϵ-close to uniform on average. Since while computing the average conditional min-entropy, the order of fixing does not matter, we can as well think of first fixing of $\mathbf{X}'_h$ and then fixing the r.v's $\{\mathbf{S}_i : i \in [h-1]\}, \{\mathbf{S}'_i : i \in [h-1]\}, \{\mathbf{R}_i : i \in [h-2]\}, \{\mathbf{R}'_i : i \in [h-2]\}$. Thus, it follows that the average conditional min-entropy of $\mathbf{X}_h$ is at least $m - 2(h-1)d_1 - \log(1/\epsilon)$.

We now show that even after fixing the r.v's $\mathbf{X}'_h, \mathbf{R}_{h-1}, \mathbf{R}'_{h-1}$, the r.v $\mathbf{S}_h$ is $2(h-1)\epsilon$-close to uniform on average. Fix $\mathbf{X}'_h$ and by the above argument $\mathbf{X}_h$ has average conditional min-entropy at least $m - 2(h-1)d_1 - \log(1/\epsilon)$. Since $\mathbf{S}_h = \mathrm{Ext}_1(\mathbf{X}_h, \mathbf{R}_{h-1})$, it follows that $\mathbf{S}_h$ is $2(h-1)\epsilon$-close to uniform on average even conditioned on $\mathbf{R}_{h-1}$. We fix $\mathbf{R}_{h-1}$, and thus $\mathbf{S}_h$ is a deterministic function of $\mathbf{X}_h$. Note that $\mathbf{S}'_h = \mathrm{Ext}_1(\mathbf{X}'_h, \mathbf{R}'_{h-1})$ is now a deterministic function of $\mathbf{R}'_h$ (and thus $\mathbf{Y}'$). Thus, we can fix $\mathbf{R}'_h$ (which also fixes $\mathbf{S}'_h$) without affecting the distribution of $\mathbf{S}_h$.

Observe that after the r.v's $\mathbf{R}_{h-1}, \mathbf{R}'_{h-1}$ are fixed, $\mathbf{S}'_h$ is a deterministic function of $\mathbf{X}'_h$. We only fix $\mathbf{S}'_h$ and do not fix $\mathbf{X}'_h$, and note that $\mathbf{S}_h$ is still $2(h-1)\epsilon$-close to uniform. Further after these fixings, each $\mathbf{X}_i$ has average conditional min-entropy at least $m - 2hd_1 - \log(1/\epsilon)$, and $\mathbf{Y}$ has average conditional min-entropy at least $k_1 - 2hd_1 - \log(1/\epsilon)$. □

By our construction of NIPM, Theorem 3.5.3 is direct from the following claim.

Claim 3.5.6. *For any $j \in [h, \ell]$, conditioned on the r.v's $\{\mathbf{S}_i : i \in [j-1]\}, \{\mathbf{S}'_i : i \in [j]\}, \{\mathbf{R}_i : i \in [j-1]\}, \{\mathbf{R}'_i : i \in [j]\}$ the following hold:*

- *$\mathbf{S}_j$ is $2(j-1)\epsilon$-close to $\mathbf{U}_d$,*
- *$\mathbf{S}_j$ is a deterministic function of $\mathbf{X}_j$*
- *for each $i \in [\ell]$, $\mathbf{X}_i$ has average conditional min-entropy at least $m - 2jd_1 - \log(1/\epsilon)$,*
- *$\mathbf{Y}$ has average conditional min-entropy at least $k_1 - 2jd_1 - \log(1/\epsilon)$,*
- *$\{\mathbf{X}_1, \ldots, \mathbf{X}_\ell, \mathbf{X}'_1, \ldots, \mathbf{X}'_\ell\}$ is independent of $\{\mathbf{Y}, \mathbf{Y}'\}$.*

Further, conditioned on the r.v's $\{\mathbf{S}_i : i \in [j]\}, \{\mathbf{S}'_i : i \in [j+1]\}, \{\mathbf{R}_i : i \in [j-1]\}, \{\mathbf{R}'_i : i \in [j]\}$ the following hold:

- *$\mathbf{R}_j$ is $(2j-1)\epsilon$-close to $\mathbf{U}_d$,*
- *$\mathbf{R}_j$ is a deterministic function of $\mathbf{Y}$,*
- *for any $i \in [\ell]$, $\mathbf{X}_i$ has average conditional min-entropy at least $m - 2(j+1)d_1 - \log(1/\epsilon)$,*
- *$\mathbf{Y}$ has average conditional min-entropy at least $k_1 - 2(j+1)d_1 - \log(1/\epsilon)$,*
- *$\{\mathbf{X}_1, \ldots, \mathbf{X}_\ell, \mathbf{X}'_1, \ldots, \mathbf{X}'_\ell\}$ is independent of $\{\mathbf{Y}, \mathbf{Y}'\}$.*

Proof. We prove this by induction on j. For the base case, when $j = h$, fix the r.v's $\{\mathbf{S}_i : i \in [h-1]\}, \{\mathbf{S}'_i : i \in [h]\}, \{\mathbf{R}_i : i \in [h-1]\}, \{\mathbf{R}'_i : i \in [h]\}$. Using Claim 3.5.5, it follows that

- $\mathbf{S}_h$ is $2(h-1)\epsilon$-close to $\mathbf{U}_d$,
- $\mathbf{S}_h$ is a deterministic function of $\mathbf{X}_h$,
- for each $i \in [\ell]$, $\mathbf{X}_i$ has average conditional min-entropy at least $m - 2hd_1 - \log(1/\epsilon)$,

- $\mathbf{Y}$ has average conditional min-entropy at least $k_1 - 2hd_1 - \log(1/\epsilon)$,
- $\{\mathbf{X}_1, \ldots, \mathbf{X}_\ell, \mathbf{X}'_1, \ldots, \mathbf{X}'_\ell\}$ is independent of $\{\mathbf{Y}, \mathbf{Y}'\}$.

Noting that $\mathbf{R}_h = Ext_2(\mathbf{Y}, \mathbf{S}_h)$, we fix $\mathbf{S}_h$ and $\mathbf{R}_h$ is $2h\epsilon$-uniform on average after this fixing. We note that $\mathbf{R}_h$ is now a deterministic function of $\mathbf{Y}$. Since $\mathbf{R}'_h$ is fixed, $\mathbf{S}'_{h+1}$ is a deterministic function of $\mathbf{X}'_{h+1}$, and we fix it without affecting the distribution of $\mathbf{R}_h$. The average conditional min-entropy of each $\mathbf{X}_i$ after these fixings is at least $m - 2(h+1)d_1 - \log(1/\epsilon)$. Further, we note that our fixings preserve the independence between $\{\mathbf{X}_1, \ldots, \mathbf{X}_\ell, \mathbf{X}'_1, \ldots, \mathbf{X}'_\ell\}$ and $\{\mathbf{Y}, \mathbf{Y}'\}$. This completes the proof of the base case.

Now suppose $j > h$. Fix the r.v's $\{\mathbf{S}_i : i \in [j-1]\}, \{\mathbf{S}'_i : i \in [j]\}, \{\mathbf{R}_i : i \in [j-2]\}, \{\mathbf{R}'_i : i \in [j-1]\}$. Using inductive hypothesis, it follows that

- $\mathbf{R}_{j-1}$ is $(2j-3)\epsilon$-close to $\mathbf{U}_d$,
- $\mathbf{R}_{j-1}$ is a deterministic function of $\mathbf{Y}$,
- for any $i \in [t]$, $\mathbf{X}_i$ has average conditional min-entropy at least $m - 2jd_1 - \log(1/\epsilon)$,
- $\mathbf{Y}$ has average conditional min-entropy at least $k_1 - 2jd_1 - \log(1/\epsilon)$,
- $\{\mathbf{X}_1, \ldots, \mathbf{X}_\ell, \mathbf{X}'_1, \ldots, \mathbf{X}'_\ell\}$ is independent of $\{\mathbf{Y}, \mathbf{Y}'\}$.

Using the fact that $\mathbf{S}_j = \mathrm{Ext}_1(\mathbf{X}_j, \mathbf{R}_{j-1})$, we fix $\mathbf{R}_{j-1}$ and $\mathbf{S}_j$ is $(2j-2)\epsilon$-close to uniform on average after this fixing. Further, $\mathbf{S}_j$ is a deterministic function of $\mathbf{X}_j$. Since $\mathbf{S}'_j$ is fixed, it follows that $\mathbf{R}'_j$ is a deterministic function of $\mathbf{Y}$ and we fix it without affecting the distribution of $\mathbf{S}_j$. We note that after these fixings, $\mathbf{Y}$ has average conditional min-entropy at least $k_1 - 2(j+1)d_1 - \log(1/\epsilon)$. Further, we note that our fixings preserve the independence between $\{\mathbf{X}_1, \ldots, \mathbf{X}_\ell, \mathbf{X}'_1, \ldots, \mathbf{X}'_\ell\}$ and $\{\mathbf{Y}, \mathbf{Y}'\}$.

Now, we fix $\mathbf{S}_j$ and it follows that $\mathbf{R}_j$ is a deterministic function of $\mathbf{Y}$ and is $(2j-1)\epsilon$-close to uniform on average. Further, since $\mathbf{R}'_j$ is fixed, it follows that $\mathbf{S}'_{j+1}$ is a deterministic function of $\mathbf{X}_{j+1}$ and we fix it without affecting the distribution of $\mathbf{R}_j$. The average conditional min-entropy

of each $\mathbf{X}_i$ after these fixings is at least $m - 2(j+1)d_1 - \log(1/\epsilon)$. Further, we note that our fixings preserve the independence between $\{\mathbf{X}_1, \ldots, \mathbf{X}_\ell, \mathbf{X}'_1, \ldots, \mathbf{X}'_\ell\}$ and $\{\mathbf{Y}, \mathbf{Y}'\}$.

This completes the proof of inductive step, and hence the claim follows. □

□

3.5.2 (ℓ, t)-Non-Malleable Independence Preserving Merger

In this section, we generalize the construction of NIPM from Section 3.5 to handle multiple adversaries.

We first introduce some notation. For a random variable $\mathbf{V}$ supported on $a \times b$ matrices, we use $\mathbf{V}_i$ to denote the random variable corresponding to the i'th row of $\mathbf{V}$. Our main result in this section is the following theorem.

Theorem 3.5.7. *There exists constant $c_{3.5.7}, c'_{3.5.7} > 0$ such that for all integers $m, d, k_1, \ell, t > 0$ and any $\epsilon > 0$, with $m \geq d \geq k_1 > c_{3.5.7}(t+1)\ell \log(m/\epsilon)$, there exists an explicit function t-NIPM $: \{0,1\}^{m\ell} \times \{0,1\}^d \to \{0,1\}^{m_1}$, $m_1 = \frac{0.9}{t}(m - c_{3.5.7}(t+1)\ell \log(m/\epsilon))$ such that if the following conditions hold:*

- *$\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t$ are r.v's, each supported on boolean $\ell \times m$ matrices s.t for any $i \in [\ell]$, $|\mathbf{X}_i - \mathbf{U}_m| \leq \epsilon$,*
- *$\{\mathbf{Y}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$ is independent of $\{\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t\}$, s.t $\mathbf{Y}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t$ are each supported on $\{0,1\}^d$ and $H_\infty(\mathbf{Y}) \geq k_1$.*
- *there exists an $h \in [\ell]$ such that $|(\mathbf{X}_h, \mathbf{X}^1_h, \ldots, \mathbf{X}^t_h) - (\mathbf{U}_m, \mathbf{X}^1_h, \ldots, \mathbf{X}^t_h)| \leq \epsilon$,*

then

$$|(\ell,t)\text{-NIPM}((\mathbf{X},\mathbf{Y}),(\ell,t)\text{-NIPM}(\mathbf{X}^1,\mathbf{Y}^1),\ldots,(\ell,t)\text{-NIPM}(\mathbf{X}^t,\mathbf{Y}^t),\mathbf{Y},\mathbf{Y}^1,\ldots,\mathbf{Y}^t$$
$$-\mathbf{U}_{m_1},(\ell,t)\text{-NIPM}(\mathbf{X}^1,\mathbf{Y}^1),\ldots,(\ell,t)\text{-NIPM}(\mathbf{X}^t,\mathbf{Y}^t),\mathbf{Y},\mathbf{Y}^1,\ldots,\mathbf{Y}^t| \leq c'_{3.5.7}\ell\epsilon.$$

Proof. We instantiate the ℓ-look-ahead extractor described in Section 3.5.1 with the following strong-seeded extractors: Let $\mathrm{Ext}_1 : \{0,1\}^m \times \{0,1\}^{d_1} \to \{0,1\}^{d_1}$, $\mathrm{Ext}_2 : \{0,1\}^d \times \{0,1\}^{d_1} \to \{0,1\}^{d_1}$ and $\mathrm{Ext}_3 : \{0,1\}^m \times \{0,1\}^{d_1} \to \{0,1\}^{m_1}$ be explicit strong-seeded from Theorem 2.1.2 designed to extract from min-entropy $k_1 = m/2, k_2 = d/2, k_3 = m - c_{3.5.7}(t+1)\log(m/\epsilon)$ respectively with error ϵ. Thus $d_1 = c_{2.1.2}\log(m/\epsilon)$.

The proof that this construction works is similar to the proof of Theorem 3.5.3, and we omit it. □

Chapter 4

Seeded Non-Malleable Extractors and Privacy Amplification

[1] Seeded non-malleable extractors were introduced by Dodis and Wichs [DW09] as a generalization of strong-seeded extractors. Recall that a (k, ϵ)-strong seeded extractor $\text{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ satisfies the property that for any (n,k)-source $\mathbf{X}$ and a typical seed s, we have $\text{Ext}(\mathbf{X}, s) \approx \mathbf{U}_m$. Informally, a non-malleable extractor nmExt satisfies the property that for a typical pair of distinct seeds (s_1, s_2), we have $\text{nmExt}(\mathbf{X}, s_1), \text{nmExt}(\mathbf{X}, s_2) \approx \mathbf{U}_{2m}$. Another way of viewing this property is the following: Fix a tampering function $\mathcal{A} : \{0,1\}^d \to \{0,1\}^d$ such that $\mathcal{A}$ has no fixed points, i.e., $\mathcal{A}(y) \neq y$ for all y. Then, a non-malleable extractor satisfies the property that for a typical seed s, the r.v $\text{nmExt}(\mathbf{X}, s)$ is close to uniform even conditioned on $\text{nmExt}(\mathbf{X}, \mathcal{A}(s))$. We now present a formal definition.

Definition 4.0.1 (Non-malleable extractor)**.** *A function* $\text{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is a* (k, ϵ)*-non-malleable extractor if the following holds: For any* (n,k)*-source* $\mathbf{X}$*, an independent*

[1] parts of this chapter have been previously published [CGL16, CL16a]

uniform seed $\mathbf{Y}$ *on* d *bits and any function* $\mathcal{A} : \{0,1\}^d \to \{0,1\}^d$ *with no fixed points,*

$$|(\text{nmExt}(\mathbf{X}, \mathbf{Y}), \text{nmExt}(X, \mathcal{A}(\mathbf{Y})), \mathbf{Y}) - (\mathbf{U}_m, \text{nmExt}(\mathbf{X}, \mathcal{A}(\mathbf{Y})), \mathbf{Y})| \leq \epsilon.$$

This generalization of a seeded extractor to satisfy this 'pairwise independence' property is non-trivial. For example, it is easy to prove that the innner product function $\text{IP} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ is not a non-malleable extractor even for min-entropy $n-1$. Recall that $\text{IP}(x,y) = \sum_i x_i y_i$ (where the sum is mod 2). By Lemma 2.5.3, it follows that IP is a 2-source extractor for min-entropy $> n/2$. Now suppose $\mathbf{X}$ is a source with its first bit fixed to 1 and each of the other $n-1$ bits are uniform and independent. Clearly $\mathbf{X}$ is an $(n, n-1)$-source. Let $\mathbf{Y}$ be an independent uniform seed. It is easy to see that if $\mathcal{A}(y)$ is the string obtained by just inverting the first bit of y (and not changing the remaining bits), then for any y, we have $\text{IP}(\mathbf{X}, y) + \text{IP}(\mathbf{X}, \mathcal{A}(y)) = 1$, implying that $\text{IP}(\mathbf{X}, y)$ fixes the value of $\text{IP}(\mathbf{X}, \mathcal{A}(y))$.

Applications to Privacy Amplification The initial motivation for non-malleable extractors comes from the problem of privacy amplification with an active adversary [BBR88,Mau92,BBCM95]. As a basic problem in information theoretic cryptography, privacy amplification deals with the case where two parties want to communicate with each other to convert their shared secret weak random source $\mathbf{X}$ into shared secret nearly uniform random bits. On the other hand, the communication channel is watched by an adversary Eve, who has unlimited computational power. To make this task possible, we assume two parties have local (non-shared) uniform random bits.

If Eve is passive (i.e., can only see the messages but cannot change them), this problem can be solved easily by applying using strong seeded extractors. However, in the case where Eve is active (i.e., can arbitrarily change, delete and reorder messages), the problem becomes much more complicated. The major challenge here is to design a protocol that uses as few number of interactions as possible, and outputs a uniform random string $\mathbf{R}$ that has length as close to $H_\infty(\mathbf{X})$ as possible (the difference is called *entropy loss*). A bit more formally, we pick a security parameter s, and if the adversary Eve remains passive during the protocol then the two parties should achieve

shared secret random bits that are 2^{-s}-close to uniform. On the other hand, if Eve is active, then the probability that Eve can successfully make the two parties output two different strings without being detected should be at most 2^{-s}.

The results in this chapter are based on joint works with Vipul Goyal and Xin Li [CGL16, CL16a].

4.1 Prior Work and Our Results in [CGL16]

There has been a long line of work on the problem of privacy amplification [MW97, DKRS06, DW09, RW03, KR09, CKOR10, DLWZ14, CRS14, Li12a, Li12b, Li15d, ADJ+14]. When the entropy rate of $\mathbf{X}$ is large, i.e., bigger than 1/2, there are known protocols that take only one round (e.g., [MW97,DKRS06]). However these protocols all have very large entropy loss. When the entropy rate of $\mathbf{X}$ is smaller than 1/2, Dodis and Wichs showed that no one round protocol exists; furthermore the length of $\mathbf{R}$ has to be at least $O(s)$ smaller than $H_\infty(\mathbf{X})$. Thus, the natural goal is to design a two-round protocol with such optimal entropy loss. However, all protocols before the work of [DLWZ14] either need to use $O(s)$ rounds, or need to incur an entropy loss of $O(s^2)$. In [DW09], Dodis and Wichs showed that explicit constructions of the non-malleable extractors can be used to give two-round privacy amplification protocols with optimal entropy loss. Using the probabilistic method, they also showed that non-malleable extractors exist when $k > 2m + 2\log(1/\varepsilon) + \log d + 6$ and $d > \log(n - k + 1) + 2\log(1/\varepsilon) + 5$. However, they were not able to give explicit constructions even for min-entropy $k = n - 1$. The first explicit construction of non-malleable extractors appeared in [DLWZ14], with subsequent improvements in [CRS14, Li12a, DY13, Li12b, ADJ+14]. All these constructions require the min-entropy of the weak source to be bigger than $0.49n$, and thus only give two-round privacy amplification protocols with optimal entropy loss for such min-entropy. Together with some other ideas, Dodis et al. also gives poly$(1/\delta)$ round protocols with optimal entropy loss for min-entropy $k \geq \delta n$, any constant $\delta > 0$. This was subsequently improved by Li [Li12b] to obtain a two-round protocol with optimal entropy loss for min-entropy $k \geq \delta n$, any constant $\delta > 0$.

In the general case, using a relaxation of non-malleable extractors called non-malleable condensers, one of the authors [Li15d] also obtained a two-round protocol with optimal entropy loss for min-entropy $k \geq C \log^2 n$, some constant $C > 1$, as long as the security parameter s satisfies $k \geq Cs^2$. For larger security parameter, the best known protocol with optimal entropy loss in [Li12b] still takes $O(s/\sqrt{k})$ rounds.

In joint work with Goyal and Li [CGL16], we construct explicit non-malleable extractors with error ε, for min-entropy $k = \Omega(\log^2(n/\epsilon))$ and seed-length $d = O(\log^2(n/\epsilon))$. In fact our construction is more general and gives explicit t-non-malleable extractors (introduced in [CRS14]), which are defined as follows.

Definition 4.1.1 (t-Non-malleable Extractor)**.** *A function* t-nmExt : $\{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is a seeded* t*-non-malleable extractor for min-entropy* k *and error* ϵ *if the following holds : If* $\mathbf{X}$ *is an* (n,k)*-source on and* $\mathcal{A}_1 : \{0,1\}^n \to \{0,1\}^n, \ldots, \mathcal{A}_t : \{0,1\}^n \to \{0,1\}^n$ *are arbitrary tampering function with no fixed points, then*

$$|t\text{-nmExt}(\mathbf{X}, \mathbf{U}_d), t\text{-nmExt}(\mathbf{X}, \mathcal{A}_1(\mathbf{U}_d)), \ldots, t - \text{-nmExt}(\mathbf{X}, \mathcal{A}_t(\mathbf{U}_d)), \mathbf{U}_d$$
$$-\mathbf{U}_m \circ t\text{-nmExt}(\mathbf{X}, \mathcal{A}_1(\mathbf{U}_d)), \ldots, t - \text{-nmExt}(\mathbf{X}, \mathcal{A}_t(\mathbf{U}_d)), U_d| < \epsilon$$

We will see in Chapter 6 that these t-non-malleable extractors are a crucial component in constructing 2-source extractors.

Theorem 1. *There exists a constant* c *such that for all* $n > 0$ *and* $\epsilon > 0$*, and* $k \geq ct \log^2\left(\frac{n}{\epsilon}\right)$*, there exists an explicit construction of a seeded* t*-non-malleable extractor* snmExt : $\{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$*, with* $m = \Omega(k/t)$ *and* $d = O(t^2 \log^2(n/\epsilon))$*.*

Combining the above theorem (with $t = 1$) with the protocol developed in [DW09], this immediately gives the following result about privacy amplification, which matches the best known result in [Li15d] but has a simpler protocol.

Theorem 2. *There exists a constant* C *such that for any* $\varepsilon > 0$ *with* $k \geq C(\log n + \log(1/\varepsilon))^2$*,*

there exists an explicit 2-round privacy amplification protocol with an active adversary for (n, k) *sources, with security parameter* $\log(1/\varepsilon)$ *and entropy loss* $O(\log n + \log(1/\varepsilon))$.

4.2 Subsequent Work and Our Results in [CL16a]

Subsequently, Cohen [Coh16a] improved our result, and constructed non-malleable extractors with seed length $d = O(\log(n/\epsilon)\log((\log n)/\epsilon))$ and min-entropy $k = \Omega(\log(n/\epsilon)\log((\log n)/\epsilon))$. In this work, he also gave another construction that worked for $k = n/(\log n)^{O(1)}$ with seed-length $O(\log n)$. In a follow up, Cohen [Coh16b] constructed non-malleable extractors with seed length $d = O(\log n + \log^3(1/\epsilon))$ and min-entropy $k = \Omega(d)$. However, in terms of the general error parameter ε, all of these results require min-entropy and seed length at least $\log^2(1/\varepsilon)$, thus none of them can be used to improve the privacy amplification protocols in [Li15d]. A recent work by Aggarwal, Hosseini and Lovett [AHL15] obtained some conditional results. In particular, they used a weaker variant of non-malleable extractors to construct privacy amplification protocols with optimal entropy loss for $k = \Omega(\log(1/\epsilon)\log n)$ assuming a conjecture in additive combinatorics.

Our first result is a new construction of non-malleable extractors that breaks the $\log^2(1/\varepsilon)$ barrier for min-entropy and seed length. Specifically, we have the following theorem.

Theorem 3. *There exists a constant* $C > 0$ *s.t for all* $n, k \in \mathbb{N}$ *and any* $\epsilon > 0$*, with* $k \geq \log(n/\epsilon)2^{C\sqrt{\log\log(n/\epsilon)}}$*, there exists an explicit* (k, ϵ)*-non-malleable extractor* $\mathrm{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$*, where* $d = \log(n/\epsilon)2^{C\sqrt{\log\log(n/\epsilon)}}$ *and* $m = k/2^{\sqrt{\log\log(n/\epsilon)}}$.

We also construct a non-malleable extractor with seed-length $O(\log n)$ for min-entropy $k = \Omega(\log n)$ and $\epsilon \geq 2^{-\log^{1-\beta}(n)}$ for any $\beta > 0$. Prior to this, explicit non-malleable extractors with seed-length $O(\log n)$ either requires min-entropy at least $n/\mathrm{poly}(\log n)$ [Coh16a] or requires $\epsilon \geq 2^{-\log^{1/3}(n)}$ [Coh16b].

Theorem 4. *There exists a constant* $C > 0$ *s.t for and all* $n, k \in \mathbb{N}$ *with* $k \geq C\log n$*, any constant* $0 < \beta < 1$*, and any* $\epsilon \geq 2^{-\log^{1-\beta}(n)}$*, there exists an explicit* (k, ϵ)*-non-malleable extractor* $\mathrm{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$*, where* $d = O(\log n)$ *and* $m = \Omega(\log(1/\epsilon))$.

Remark 4.2.1. *A careful examination reveals that our seed length and min-entropy requirement are better than those of [Coh16a, Coh16b] in all cases except the case that ε is large enough (e.g., $\epsilon \geq 2^{-\log^{1/3}(n)}$), where both [Coh16b] and our results require seed length and min-entropy $O(\log n)$.*

Note that given any error parameter ε, our non-malleable extractor in Theorem 3 only requires min-entropy and seed length $\log^{1+o(1)}(n/\varepsilon)$.

We also show how to further lower the min-entropy requirement of the non-malleable extractor in Theorem 3 at the expense of using a larger seed. We complement this result by constructing another non-malleable extractor with shorter seed-length than in Theorem 3 at the expense of larger entropy. We now state these results more formally.

Theorem 5. *For all $n, k \in \mathbb{N}$ and any $\epsilon > 0$, with $k \geq \log(n/\epsilon)2^{2^{\Omega(\sqrt{\log\log\log(n/\epsilon)})}}$, there exists an explicit (k, ϵ)-non-malleable extractor* $\mathrm{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$*, where $d = (\log(n/\epsilon))^3 2^{(\log\log\log(n/\epsilon))^{O(1)}}, m = \Omega(k)$.*

Theorem 6. *For all $n, k \in \mathbb{N}$ and any $\epsilon > 0$, with $k \geq (\log(n/\epsilon))^3 2^{(\log\log\log(n/\epsilon))^{O(1)}}$, there exists an explicit (k, ϵ)-non-malleable extractor* $\mathrm{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$*, where $d = \log(n/\epsilon)2^{2^{O(\sqrt{\log\log\log(n/\epsilon)})}}, m = \frac{k}{\log(n/\epsilon)2^{(\log\log\log(n/\epsilon))^{O(1)}}} - O((\log(n/\epsilon))^2)$.*

Privacy Amplification Using Theorem 3 and the protocol in [DW09], we immediately obtain a two-round privacy amplification protocol with optimal entropy loss, for almost all possible security parameters.

Theorem 7. *There exists a constant $C > 0$ such that for any security parameter s with $k \geq (s + \log n)2^{C\sqrt{\log(s+\log n)}}$, there exists an explicit 2-round privacy amplification protocol for (n, k)-sources with entropy loss $O(\log n + s)$ and communication complexity $(s + \log n)2^{O(\sqrt{\log(s+\log n)})}$, in the presence of an active adversary.*

In particular, this gives us two-round privacy amplification protocols with optimal entropy loss for security parameter $s \leq k^{1-\alpha}$ for any constant $\alpha > 0$.

Reference	Min-Entropy	Seed Length
[DW09] (non-constructive)	$> 2m + 2\log(1/\epsilon) + \log d + 6$	$> \log(n-k+1)+2\log(1/\epsilon)+5$
[DLWZ14]	$> n/2$	n
[CRS14, Li12a, DY13]	$> n/2$	$O(\log(n/\epsilon))$
[Li12b]	$0.49n$	n
Theorem 1	$\Omega((\log(n/\epsilon))^2)$	$O((\log(n/\epsilon))^2)$
[Coh16a]	$\Omega(\log(n/\epsilon)\log((\log n)/\epsilon))$	$O(\log(n/\epsilon)\log((\log n)/\epsilon))$
[Coh16b]	$\Omega(\log n + (\log(1/\epsilon))^3)$	$O(\log n + (\log(1/\epsilon))^3)$
Theorem 3	$\log(n/\epsilon)2^{\Omega(\sqrt{\log\log(n/\epsilon)})}$	$\log(n/\epsilon)2^{\Omega(\sqrt{\log\log(n/\epsilon)})}$
Theorem 6	$\log(n/\epsilon)2^{2^{\Omega(\sqrt{\log\log\log(n/\epsilon)})}}$	$(\log(n/\epsilon))^{3+o(1)}$
Theorem 5	$(\log(n/\epsilon))^{3+o(1)}$	$\log(n/\epsilon)2^{2^{O(\sqrt{\log\log\log(n/\epsilon)})}}$

Table 4.1: A summary of results on non-malleable extractors

Instead if we use the non-malleable extractor from Theorem 5, we obtain a two-round privacy amplification protocol with optimal entropy loss, for even smaller min-entropy (at the expense of larger communication complexity). More formally, we have the following theorem.

Theorem 8. *There exists a constant $C > 0$ such that for any security parameter s with $k \geq (s+\log n)2^{2^{C\sqrt{\log\log(s+\log n)}}}$, there exists an explicit 2-round privacy amplification protocol for (n,k)-sources with entropy loss $O(\log n + s)$ and communication complexity $(s+\log n)^3 2^{(\log\log(s+\log n))^{O(1)}}$, in the presence of an active adversary.*

4.3 A Non-Malleable Extractor for $\log^2(n/\epsilon)$ min-entropy

In section, we present the construction of a seeded t-non-malleable extractor that works for min-entropy $k = \Omega(t\log^2(n/\epsilon))$ and requires seed-length $d = O(t^2\log^2(n/\epsilon))$. A key ingredient in this

construction is an explicit correlation breaker with advice constructed in Chapter 3. We first set up the various ingredients in the construction with appropriate parameters.

Subroutines and Parameters

1. Let t be a parameter.

2. Let $n_1 = \log\left(\frac{tn}{\epsilon}\right)$. Let $\text{Ext}_s : \{0,1\}^n \times \{0,1\}^{n_1} \to \{0,1\}^{n_1}$ be the strong seeded extractor from Theorem 2.1.2 set to extract from min-entropy $2n_1$ and error $2^{-\Omega(n_1)}$.

3. Let $\mathcal{C}$ be an explicit $[\frac{d}{\alpha}, d, \frac{1}{10}]$-binary linear error correcting code with encoder $E : \{0,1\}^d \to \{0,1\}^{\frac{d}{\alpha}}$. Such explicit codes are known, for example from the work of Alon et al. [ABN+92].

4. Let $\text{Ext}_{\text{Samp}} : \{0,1\}^{n_1} \times \{0,1\}^{d_1} \to \{0,1\}^{n_2}$ be the strong seeded extractor from Theorem 2.1.4 set to extract from min-entropy $\frac{n_1}{2}$ with error $\frac{1}{20}$ and output length n_2, such that $N_2 D_1 = \frac{d}{\alpha}$, where $N_2 = 2^{n_2}$ and $D_1 = 2^{d_1}$. Let $\{0,1\}^{d_1} = \{s_1, \ldots, s_{D_1}\}$. Define $\text{Samp} : \{0,1\}^{n_1} \to [\frac{d}{\alpha}]^{D_1}$ as: $\text{Samp}(x) = (\text{Ext}(x, s_1) \circ s_1, \ldots, \text{Ext}(x, s_{D_1}) \circ s_{D_1})$. By Theorem 2.1.4, we have $D_1 = c_1 n_1$, for some constant c_1.

5. Let $\ell = n_1 + D_1 = (c_1 + 1)n_1$.

6. We set up the parameters for the components used by flip-flop (computed by Algorithm 1) as follows.

   (a) Let $n_3 = c_3 t\ell, n_4 = 10\ell$, for some large enough constant c_3.

   Let $\text{Ext}_q : \{0,1\}^{n_3} \times \{0,1\}^{n_4} \to \{0,1\}^{n_4}$ be the strong seeded extractor from Theorem 2.1.2 set to extract from min-entropy $k_q = \frac{n_3}{4}$ with error $\epsilon = 2^{-\Omega(n_4)}$.

   Let $\text{Ext}_w : \{0,1\}^n \times \{0,1\}^{n_4} \to \{0,1\}^{n_4}$ be the strong seeded extractor from Theorem 2.1.2 set to extract from min-entropy $\frac{k}{2}$ with error $\epsilon = 2^{-\Omega(n_4)}$.

   (b) Let $\text{laExt} : \{0,1\}^n \times \{0,1\}^{n_3+n_4} \to \{0,1\}^{2n_4}$ be the look ahead extractor used by 2laExt. Recall that the parameters in the alternating extraction protocol are set as $m = n_4, u = 2$ where u is the number of steps in the protocol, m is the length of each random

variable that is communicated between the players, and $\text{Ext}_q, \text{Ext}_w$ are the strong seeded extractors used in the protocol.

(c) Let $\text{Ext} : \{0,1\}^d \times \{0,1\}^{n_4} \to \{0,1\}^{n_3}$ be the strong seeded extractor from Theorem 2.1.2 set to extract from min-entropy $\frac{d}{2}$ with seed length n_4 and error $2^{-\Omega(n_4)}$.

7. Let nmExt_1 be the function computed by Algorithm 2, which uses the function 2laExt set up as above.

8. Let $n_5 = \frac{k}{100t}$. Let $\text{Ext}_1 : \{0,1\}^n \times \{0,1\}^{n_4} \to \{0,1\}^{n_5}$ be the strong seeded extractor from Theorem 2.1.2 set to extract from min-entropy $\frac{k}{4}$ with seed length n_4, error $2^{-\Omega(n_4)}$.

Algorithm 5: snmExt(x,y)
Input: Bit strings x, y, of length n, d respectively.
Output: A bit string of length n_4.

1 $y_1 = \text{Slice}(y, n_1)$. Compute $v = \text{Ext}_s(x, y_1)$.
2 Compute $T = \text{Samp}(v) \subset [\frac{n}{\alpha}]$.
3 Let $z = y_1 \circ y_2$ where $y_2 = (E(y))_{\{T\}}$.
4 Output $\text{Ext}_1(x, \text{nmExt}_1(x, y, z))$.

We now state our main theorem.

Theorem 4.3.1. *Let* $\text{snmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^{n_5}$ *be the function computed by Algorithm 4. Then* snmExt *satisfies the following property: For any* $\epsilon > 0$, $k \geq \Omega(t \log^{2+\gamma}(\frac{n}{\epsilon}))$, *and* $d = O(t^2 \log^2(\frac{n}{\epsilon}))$, *if* $\mathbf{X}$ *is a* (n,k)*-source, and* $\mathbf{Y}$ *is an independent and uniform distribution on* $\{0,1\}^d$, *and* $\mathcal{A}_1 \ldots, \mathcal{A}_t$ *are arbitrary tampering functions, such that for each* $i \in [t]$, $\mathcal{A}_i$ *has no fixed points, then the following holds:*

$$|\text{snmExt}(\mathbf{X}, \mathbf{Y}), \text{snmExt}(\mathbf{X}, \mathcal{A}_1(\mathbf{Y})), \ldots, \text{snmExt}(\mathbf{X}, \mathcal{A}_t(\mathbf{Y})), \mathbf{Y} -$$
$$\mathbf{U}_{n_5}, \text{snmExt}(\mathbf{X}, \mathcal{A}_1(\mathbf{Y})), \ldots, \text{snmExt}(\mathbf{X}, \mathcal{A}_t(\mathbf{Y})), \mathbf{Y}| \leq O(\epsilon),$$

Notation: For any function H, if $\mathbf{V} = H(\mathbf{X}, \mathbf{Y})$, let $\mathbf{V}^i$ denote the random variable $H(\mathbf{X}, \mathcal{A}_i(\mathbf{Y}))$.

Proof. We first prove the following claim.

Claim 4.3.2. *With probability at least* $1-\epsilon$, $\mathbf{Z} \neq \mathbf{Z}^i$ *for each* $i \in [t]$.

Proof. Pick an arbitrary $i \in [t]$. If $\mathbf{Y}_1 \neq \mathbf{Y}_1^i$, then we have $\mathbf{Z} \neq \mathbf{Z}^i$. Now suppose $\mathbf{Y}_1 = \mathbf{Y}_1^i$. We fix $\mathbf{Y}_1$, and note that since Ext_s is a strong extractor (Theorem 2.5.3), B is $2^{-\Omega(n_1)}$-close to $\mathbf{U}_{n_1}$.

Since $\mathcal{A}_i$ has no fixed points, it follows that since E is an encoder of a code with relative distance distance $\frac{1}{10}$, $\Delta(E(\mathbf{Y}), E(\mathbf{Y}^i)) \geq \frac{d}{10\alpha}$. Let $\mathcal{D} = \{j \in \left[\frac{d}{\alpha}\right] : E(\mathbf{Y})_{\{j\}} \neq E(\mathbf{Y}^i)_{\{j\}}\}$. Thus $|\mathcal{D}| \geq \frac{d}{10\alpha}$. Using Theorem 2.4.2, it follows that with probability at least $1-\epsilon$, $|\mathcal{D} \cap \mathrm{Samp}(\mathbf{V})| \geq 1$, and thus $\mathbf{Y}_2 \neq \mathbf{Y}_2^{(i)}$ (since $\mathrm{Samp}(\mathbf{V}) = \mathrm{Samp}(\mathbf{V}^i)$). The claim now follows by a simple union bound. □

We fix $\mathbf{Z}, \mathbf{Z}^1, \ldots, \mathbf{Z}^t$ such that $\mathbf{Z} \neq \mathbf{Z}^i$ for any $i \in [t]$ (from the lemma above, this occurs with probability $1-\epsilon$). We note that by the Lemma 4.3.2 and Lemma 2.3.7, the source $\mathbf{X}$ has min-entropy at least $k - 2n_1$ and the source $\mathbf{Y}$ has min-entropy at least $d - 2\ell$ with probability at least $1-\epsilon$.

Lemma 4.3.1 now follows directly from Lemma 3.3.1 by noting that the following hold by our choice of parameters:

- $\frac{d}{2} > 20\ell(t(n_3+n_4) + \log(\frac{1}{\epsilon}))$
- $k - 2n_1 \geq \frac{n_3}{4} + 20\ell(tn_4 + \log(\frac{1}{\epsilon}))$
- $n_3 - 2n_1 \geq \frac{4}{3}(10tn_4 + 2\log(\frac{1}{\epsilon}))$

This concludes the proof. □

4.4 Near Optimal Non-Malleable Extractors

We present an explicit construction of a non-malleable extractor with min-entropy requirement $k = (\log(n/\epsilon))^{1+o(1)}$ and seed-length $d = (\log(n/\epsilon))^{1+o(1)}$. We also show a way of setting parameters

that allows for $O(\log n)$ seed-length for large enough error. The following are the main results of this section.

Theorem 4.4.1. *There exist a constant $C_{4.4.1} > 0$ s.t for all $n, k \in \mathbb{N}$ and any $\epsilon > 0$, with $k \geq \log(n/\epsilon)2^{C_{4.4.1}\sqrt{\log\log(n/\epsilon)}}$, there exists an explicit (k,ϵ)-non-malleable extractor* $\text{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$, *where $d = \log(n/\epsilon)2^{C_{4.4.1}\sqrt{\log\log(n/\epsilon)}}$ and $m = k/2^{\sqrt{\log\log(n/\epsilon)}}$.*

Theorem 4.4.2. *There exist a constant $C_{4.4.2} > 0$ s.t for constant $\beta > 0$ and all $n, k \in \mathbb{N}$ and any $\epsilon > 2^{-\log^{1-\beta}(n)}$, with $k \geq C_{4.4.2}\log n$, there exists an explicit (k,ϵ)-non-malleable extractor* $\text{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$, *where $d = O(\log n)$ and $m = \Omega(\log(1/\epsilon))$.*

We derive both the above theorems from the following theorem.

Theorem 4.4.3. *There exist constants $\delta_{4.4.3}, C_{4.4.3} > 0$ s.t for all $n, k \in \mathbb{N}$ and any error parameter $\epsilon_1 > 0$, with $k \geq \log(k/\epsilon_1)2^{C_{4.4.3}\sqrt{\log\log(n/\epsilon_1)}} + C_{4.4.3}\log(n/\epsilon_1)$, there exists an explicit (k,ϵ')-non-malleable extractor* $\text{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$, *where $d = \log(k/\epsilon)2^{C_{4.4.3}\sqrt{\log\log(n/\epsilon_1)}} + C_{4.4.3}\log(n/\epsilon_1), m = \delta_{4.4.3}k/2^{\sqrt{\log\log(n/\epsilon_1)}}$ and $\epsilon' = C_{4.4.3}\epsilon_1\log(n/\epsilon_1)$.*

We first show how to derive Theorem 4.4.1 and Theorem 4.4.2 from Theorem 4.4.3.

Proof of Theorem 4.4.1. Let $\text{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ be the function from Theorem 4.4.3 set to extract from min-entropy k, where we set the parameter $\epsilon_1 = \epsilon/2C_{4.4.3}n$. It follows that the error of nmExt is

$$C_{4.4.3}\epsilon_1\log(n/\epsilon_1) = \frac{\epsilon}{2n}(\log n + \log(2C_{4.4.3}n) + \log(1/\epsilon)) < \epsilon.$$

Further note that for this setting of ϵ_1, the min-entropy required and seed length are $\log(n/\epsilon)2^{C_{4.4.1}\sqrt{\log\log(n/\epsilon)}} + C_{4.4.1}\log(n/\epsilon)$, for some constant $C_{4.4.1}$. □

Proof of Theorem 4.4.2. Let $\text{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ be the function from Theorem 4.4.3 set to extract from min-entropy $2C_{4.4.3}\log(n/\epsilon_1)$, where we set the parameter $\epsilon_1 =$

$\epsilon/2C_{4.4.3}\log n$. Thus, the error of nmExt is

$$\epsilon_1 \log(n/\epsilon_1) \leq \frac{\epsilon}{2\log n}(\log n + \log(1/\epsilon) + \log(2C_{4.4.3}\log n)) < \epsilon.$$

For this setting of parameters, we note that the seed-length required by nmExt is bounded by $\log((\log^2 n)/\epsilon)2^{C_{4.4.3}\sqrt{\log\log(n\log n/\epsilon)}} + C_{4.4.3}\log(n\log n/\epsilon) = O(\log n)$. □

We spend the rest of the section proving Theorem 4.4.3. A key ingredient in our construction in an explicit non-malleable independence preserving merger with strong parameters.

4.4.1 A Recursive Non-Malleable Independence Preserving Merger

In this section, we show a recursive way of applying the (ℓ, t)-NIPM constructed in the previous section in order to achieve better trade-off between parameters. This object is crucial in obtaining our near optimal non-malleable extractor construction.

Notation: For an $a \times b$ matrix $\mathbf{V}$, and any $\mathbf{S} \subseteq [a]$, let $\mathbf{V_S}$ denote the matrix obtained by restricting $\mathbf{V}$ to the rows indexed by $\mathbf{S}$.

Our main result in this section is the following theorem.

Theorem 4.4.4. *For all integers $m, \ell, L, t > 0$, any $\epsilon > 0$, $r = \lceil \frac{\log L}{\log \ell} \rceil$ and any $d = (c_{3.5.7}\ell\log(m/\epsilon) + d')(t+2)^{r+1}$, there exists an explicit function (L, ℓ, t)-NIPM $: \{0,1\}^{mL} \times \{0,1\}^d \to \{0,1\}^{m'}$, $m' = (0.9/t)^r(m - c_{3.5.7}\ell(t+1)r\log(m/\epsilon))$, such that if the following conditions hold:*

- *$\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t$ are r.v's, each supported on boolean $L \times m$ matrices s.t for any $i \in [L]$, $|\mathbf{X}_i - \mathbf{U}_m| \leq \epsilon$,*
- *$\{\mathbf{Y}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$ is independent of $\{\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t\}$, s.t $\mathbf{Y}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t$ are each supported on $\{0,1\}^d$ and $H_\infty(\mathbf{Y}) \geq d - d'$,*
- *there exists an $h \in [\ell]$ such that $|(\mathbf{X}_h, \mathbf{X}^1_h, \ldots, \mathbf{X}^t_h) - (\mathbf{U}_m, \mathbf{X}^1_h, \ldots, \mathbf{X}^t_h)| \leq \epsilon$,*

then

$$|(L,\ell,t)\text{-NIPM}((\mathbf{X},\mathbf{Y}),(L,\ell,t)\text{-NIPM}(\mathbf{X}^1,\mathbf{Y}^1),\ldots,(L,\ell,t)\text{-NIPM}(\mathbf{X}^t,\mathbf{Y}^t),\mathbf{Y},\mathbf{Y}^1,\ldots,\mathbf{Y}^t$$
$$-\mathbf{U}_{m_1},(L,\ell,t)\text{-NIPM}(\mathbf{X}^1,\mathbf{Y}^1),\ldots,(L,\ell,t)\text{-NIPM}(\mathbf{X}^t,\mathbf{Y}^t),\mathbf{Y},\mathbf{Y}^1,\ldots,\mathbf{Y}^t| \leq 2c'_{3.5.7}L\epsilon.$$

Proof. We set up parameters and ingredients required in our construction.

- For $i \in [r]$, let $L_i = \lceil \frac{L}{\ell^i} \rceil$.
- Let $d_1 = d' + \log(1/\epsilon) + c_{3.5.7}(t+1)\ell\log(m/\epsilon)$. For $i \in [r]$, let $d_i = (t+2)d_{i-1}$.
- Let $m_0 = m$. For $i \in [r]$, define $m_i = 0.9^i(m - ic_{3.5.7}(t+1)\ell\log(m/\epsilon))$
- For each $i \in [r]$, let (ℓ,t)-NIPM$_i : \{0,1\}^{\ell m_i} \times \{0,1\}^{d_i} \to \{0,1\}^{m_{i+1}}$ be an instantiation of the function from Theorem 3.5.7 with error parameter ϵ.

Algorithm 6: (L,ℓ,t)-NIPM(x,y)
Input: x is a boolean $L \times m$ matrix, and y is a bit string of length d.
Output: A bit string of length m_r.

```
Let x[0] = x.
for i = 1 to r do
    Let y[i] = Slice(y, d_i)
    Let x[i] be a L_i × m_i matrix, whose j'th row
    x[i]_j = (ℓ,t)-NIPM_i(x[i−1]_{[(j−1)ℓ+1, jℓ]}, y[i])
end
Ouput x[r].
```

We prove the following claim from which it is direct that the function (L,ℓ,t)-NIPM computed by Algorithm 6 satisfies the conclusion of Theorem 4.4.4. Let $\epsilon_0 = \epsilon$, and for $i \in [r]$, let $\epsilon_i = \ell\epsilon_{i-1} + c'_{3.5.7}\ell\epsilon$.

Claim 4.4.5. *For all $i \in [r]$, conditioned on the r.v's $\{\mathbf{Y}[j] : j \in [i]\}, \{\mathbf{Y}^g[j] : j \in [i], g \in [t]\}$, the following hold:*

- $\mathbf{X}[i], \mathbf{X}^1[i], \ldots, \mathbf{X}^t[i]$ *are r.v's, each supported on boolean* $L_i \times m_i$ *matrices s.t for any* $j \in [L_i]$, $|\mathbf{X}[i]_j - \mathbf{U}_{m_i}| \leq (c'_{3.5.7}\ell)^i \epsilon$,
- $\{\mathbf{Y}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$ *is independent of* $\{\mathbf{X}[i], \mathbf{X}[i]^1, \ldots, \mathbf{X}[i]^t\}$.
- *there exists an* $h_i \in [L_i]$ *such that* $\mathbf{X}[i]_h | \{\mathbf{X}[i]^1_h, \ldots, \mathbf{X}[i]^t_h\}$ *is* ϵ_i*-close to* $\mathbf{U}_{m_i}$ *on average,*
- $\mathbf{Y}$ *has average conditional min-entropy at least* $d - d_{i+1} + c_{3.5.7}(t+1)\ell \log(m/\epsilon)$.

Proof. We prove this claim by an induction on i. The base case, when $i = 0$, is direct. Thus suppose $i \geq 1$. Fix the *r.v*'s $\{\mathbf{Y}[j] : j \in [i-1]\}, \{\mathbf{Y}^g[j] : j \in [i-1], g \in [t]\}$. By inductive hypothesis, it follows that

- $\mathbf{X}[i-1], \mathbf{X}^1[i-1], \ldots, \mathbf{X}^t[i-1]$ are r.v's each supported on boolean $L_{i-1} \times m_{i-1}$ matrices s.t for any $j \in [L_{i-1}]$, $|\mathbf{X}[i-1]_j - \mathbf{U}_{m-1}| \leq (c'_{3.5.7}\ell)^{i-1}\epsilon$,
- $\{\mathbf{Y}[i-1], \mathbf{Y}^1[i-1], \ldots, \mathbf{Y}^t[i-1]\}$ is independent of $\{\mathbf{X}[i-1], \mathbf{X}[i-1]^1, \ldots, \mathbf{X}[i-1]^t\}$.
- $h_i \in [L_i]$ such that $\mathbf{X}[i-1]_h | \{\mathbf{X}[i-1]^1_h, \ldots, \mathbf{X}[i-1]^t_h\}$ is ϵ_{i-1}-close to $\mathbf{U}_{m_{i-1}}$ on average,
- $\mathbf{Y}$ has average conditional min-entropy at least $d - d_i + c_{3.5.7}(t+1)\ell \log(m/\epsilon)$.

Thus the r.v $\mathbf{Y}[i] = \text{Slice}(\mathbf{Y}, d_i)$ has average conditional min-entropy at least $c_{3.5.7}(t+1)\ell \log(n/\epsilon)$. Let $h_i \in [\ell(h_i - 1) + 1, \ell h_i]$, for some $h_i \in [L_i]$. It follows that conditioned on the r.v's $\mathbf{Y}[i], \{\mathbf{Y}^g[i] : g \in [t]\}$, for any $j \in [L_i]$, $|\mathbf{X}[i]_j - \mathbf{U}_m| \leq \ell\epsilon_{i-1} + c'_{3.5.7}\ell\epsilon = \epsilon_i$.

Further, using Theorem 3.5.7, conditioned on $\mathbf{Y}[i], \{\mathbf{Y}^g[i] : g \in [t]\}, \{\mathbf{X}^g[i]_{h_i} : g \in [t]\}$, the r.v $\mathbf{X}[i]_{h_i}$ is $\ell\epsilon_{i-1} + c'_{3.5.7}\ell\epsilon$-close to uniform on average.

Thus, we fix the r.v's $\mathbf{Y}[i], \{\mathbf{Y}^g[i] : g \in [t]\}$, and note that $\mathbf{Y}$ still has average conditional min-entropy at least $d - d_i + c_{3.5.7}(t+1)\ell \log(m/\epsilon) - (t+1)d_i \geq d - d_{i+1} + c_{3.5.7}(t+1)\ell \log(m/\epsilon)$. This completes the proof of the inductive step, and the theorem follows. □

□

4.4.2 The Non-Malleable Extractor Construction

The following function is implicit in the construction in Section 4.3. Informally, advGen takes as input a source $\mathbf{X}$ and a seed $\mathbf{Y}$ and produces a short string such that for any r.v $\mathbf{Y}' \neq \mathbf{Y}$, $\text{advGen}(\mathbf{X}, \mathbf{Y}) \neq \text{advGen}(\mathbf{X}, \mathbf{Y})$. We record this property more formally.

Theorem 4.4.6. *There exists a constant $c_{4.4.6}, C_{4.4.6} > 0$ such that for all $n > 0$ and any $\epsilon > 0$, there exists an explicit function* $\text{advGen} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^L$, *$L = c_{4.4.6} \log(n/\epsilon)$ satisfying the following: Let $\mathbf{X}$ be an (n,k)-source, and $\mathbf{Y}$ be an independent uniform seed on d bits. Let $\mathbf{Y}'$ be a r.v on d bits independent of $\mathbf{X}$, s.t $\mathbf{Y}' \neq \mathbf{Y}$. If $k, d \geq C_{4.4.6} \log(n/\epsilon)$, then*

- *with probability at least $1 - \epsilon$,* $\text{advGen}(\mathbf{X}, \mathbf{Y}) \neq \text{advGen}(\mathbf{X}, \mathbf{Y}')$,
- *there exists a function f such that conditioned on* $\text{advGen}(\mathbf{X}, \mathbf{Y}), \text{advGen}(\mathbf{X}, \mathbf{Y}'), f(\mathbf{X})$,
  - $\mathbf{X}$ *remains independent of* $\mathbf{Y}, \mathbf{Y}'$,
  - $\mathbf{X}$ *has average conditional min-entropy at least* $k - C_{4.4.6} \log(n/\epsilon)$,
  - $\mathbf{Y}$ *has average conditional min-entropy at least* $d - C_{4.4.6} \log(n/\epsilon)$

We are now ready to prove Theorem 4.4.3.

Proof of Theorem 4.4.3. We set up parameters and ingredients required in our construction.

- Let $\text{advGen} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^L$, $L = c_{4.4.6} \log(n/\epsilon_1)$, be the function from Theorem 4.4.6 with error parameter ϵ_1.
- Let $d_1 = (C_{4.4.6} + C + 1) \log(n/\epsilon_1)$, for some large enough constant C.
- Let $\text{flip-flop} : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^{m'}$, $m' = \delta k$, be the function computed by Algorithm 1 with error parameter ϵ_1.
- $d_2 = c_{2.1.2} \log(d/\epsilon_1), d_3 = c_{2.1.2} \log(m'/\epsilon_1)$.

- Let $\text{Ext}_1 : \{0,1\}^n \times \{0,1\}^{d_2} \to \{0,1\}^{d'}$, $d' = 0.9d - 2d_1 - C_{4.4.6}\log(n/\epsilon_1)$ be a $(d - 2d_1 - C_{4.4.6}\log(n/\epsilon_1), \epsilon_1)$-strong-seeded extractor from Theorem 2.1.2.
- Let $\text{Ext}_2 : \{0,1\}^{m'} \times \{0,1\}^{d_3} \to \{0,1\}^{m''}$, $m'' = 0.9m' - 2d_2$, be a $(m' - 2d_2 - \log(1/\epsilon_1), \epsilon_1)$-strong-seeded extractor from Theorem 2.1.2.
- Let $\ell = 2^{\sqrt{\log L}}$.
- Let $(L, \ell, 1)$-NIPM : $\{0,1\}^{Lm''} \times \{0,1\}^{d'} \to \{0,1\}^m$ be the function from Theorem 4.4.4, $m = 0.9^r m' - 2c_{3.5.7}\ell(t+1)r\log(m/\epsilon_1)$ with error parameter ϵ_1.

Algorithm 7: nmExt(x, y)
Input: x, y are bit string of length n, d respectively.
Output: A bit string of length m.

1 Let $w = \text{advGen}(x, y)$.
2 Let $y = y_1 \circ y_2$, where $y_1 = \text{Slice}(y, d_1)$.
3 Let v be a $L \times m'$ matrix, whose i'th row $v_i = \text{flip-flop}(x, y_1, w_i)$ (w_i is the i'th bit of the string w).
4 Let $\overline{v_1} = \text{Slice}(v_1, d_2)$
5 Let $\overline{y} = \text{Ext}_1(y, \overline{v_1}) = \overline{y_1} \circ \overline{y_2}$, where $\overline{y_1} = \text{Slice}(\overline{y}, d_3)$.
6 Let z be a $L \times m''$ matrix, whose i'th row $z_i = \text{Ext}_2(v_i, \overline{y_1})$
7 Output $\overline{z} = (L, \ell, 1)\text{-NIPM}(z, \overline{y})$.

We prove in the following claims that the function nmExt constructed in Algorithm 8 satisfies the conclusion of Theorem 4.4.3. Let $\mathcal{A}$ be the adversarial function tampering the seed $\mathbf{Y}$, and let $\mathbf{Y}' = \mathcal{A}(\mathbf{Y})$. Since $\mathcal{A}$ has no fixed points, it follows that $\mathbf{Y} \neq \mathbf{Y}'$.

Notation: For any random variable $\mathbf{H} = g(\mathbf{X}, \mathbf{Y})$ (where g is an arbitrary deterministic function), let $\mathbf{H}' = g(\mathbf{X}, \mathbf{Y}')$.

Claim 4.4.7. *With probability at least* $1 - \epsilon$, $\mathbf{W} \neq \mathbf{W}'$.

Proof. Follows directly from Theorem 4.4.6. □

Let f be the function guaranteed by Theorem 4.4.6.

Claim 4.4.8. *Conditoned on the r.v's* $\mathbf{W}, \mathbf{W}', \mathbf{Y}_1, \mathbf{Y}_1', f(\mathbf{X})$*, the following hold:*

- *for each* $i \in [L]$*,* $\mathbf{V}_i$ *is* ϵ_1*-close to uniform,*
- *there exists an* $h \in [L]$ *such that conditioned on* $\mathbf{V}_h'$*, the r.v* $\mathbf{V}_h$ *is* ϵ_1*-close to uniform on average,*
- $\{\mathbf{V}, \mathbf{V}'\}$ *is independent of* $\{\mathbf{Y}, \mathbf{Y}'\}$*.*
- $\mathbf{Y}$ *has average conditional min-entropy at least* $d - C_{4.4.6}\log(n/\epsilon_1) - 2d_1$*.*

Proof. Fix the r.v's $\mathbf{W}, \mathbf{W}', f(\mathbf{X})$ such that $\mathbf{W} \neq \mathbf{W}'$. It follows from Theorem 4.4.6 that after this conditioning,

- $\mathbf{X}$ is independent of $\mathbf{Y}, \mathbf{Y}'$,
- $\mathbf{X}$ has average conditional min-entropy at least $k - C_{4.4.6}\log(n/\epsilon_1)$,
- $\mathbf{Y}$ has average conditional min-entropy at least $d - C_{4.4.6}\log(n/\epsilon_1)$

Thus $\mathbf{Y}_1 = \text{Slice}(\mathbf{Y}, d_1)$ has average conditional min-entropy at least $O(\log(n/\epsilon_1))$. The claim now follows by applying Lemma 3.2.1. □

Claim 4.4.9. *Conditioned on the r.v's* $\mathbf{W}, \mathbf{W}', \overline{\mathbf{V}}_1, \overline{\mathbf{V}_1}', \mathbf{Y}_1, \mathbf{Y}_1'\overline{\mathbf{Y}_1}, \overline{\mathbf{Y}_1}', f(\mathbf{X})$*, the following hold:*

- $\overline{\mathbf{Y}}$ *has average conditional min-entropy at least* $d' - 2d_3 - \log(1/\epsilon)$*.*
- *for each* $i \in [L]$*,* $\mathbf{Z}_i$ *is* $3\epsilon_1$*-close to uniform on average.*
- *there exists* $h \in [L]$ *such that further conditioned on* $\mathbf{Z}_i'$*,* $\mathbf{Z}_i$ *is* $3\epsilon_1$*-close to uniform on average.*
- $\{\overline{\mathbf{Y}}, \overline{\mathbf{Y}}'\}$ *is independent of* $\{\mathbf{Z}, \mathbf{Z}'\}$*.*

Proof. Fix the r.v's $\mathbf{W}, \mathbf{W}', \mathbf{Y}_1, \mathbf{Y}_1', f(\mathbf{X})$. By Claim 4.4.8, we have

- for each $i \in [L]$, $\mathbf{V}_i$ is ϵ_1-close to uniform,

- there exists an $h \in [L]$ such that conditioned on $\mathbf{V}'_h$, the r.v $\mathbf{V}_h$ is ϵ_1-close to uniform on average,
- $\{\mathbf{V}, \mathbf{V}'\}$ is independent of $\{\mathbf{Y}, \mathbf{Y}'\}$.
- $\mathbf{Y}$ has average conditional min-entropy at least $d - C_{4.4.6}\log(n/\epsilon_1) - 2d_1$.

Using the fact that Ext_1 is a strong extractor, it follows that we can fix $\overline{\mathbf{V}_1}$, and $\overline{\mathbf{Y}}$ is $2\epsilon_1$-close to uniform on average. Further, $\overline{\mathbf{Y}}$ is a deterministic function of $\mathbf{Y}$. Thus, we fix $\overline{\mathbf{V}_1}'$ without affecting the distribution of $\overline{\mathbf{Y}}$. Now, using the fact that Ext_2 is a strong extractor, we can fix $\overline{\mathbf{Y}_1}$, and we have for each $i \in [L]$, $\mathbf{Z}_i$ is $3\epsilon_1$-close to uniform on average. Next we can fix $\overline{\mathbf{Y}_1}'$ without affecting $\mathbf{V}$.

We prove that conditioned on $\mathbf{Z}'_i$, the r.v $\mathbf{Z}_i$ is $3\epsilon_1$-close to uniform on average in the following way. For this argument, as above we fix all r.v's but do not yet fix $\overline{\mathbf{Y}_1}, \overline{\mathbf{Y}_1}'$. Instead, we first fix $\mathbf{V}'_h$, and $\mathbf{V}_h$ has average conditional min-entropy at least $m' - 2d_2$. We now fix $\overline{\mathbf{Y}_1}$, and as before we have $\mathbf{Z}_h$ is $3\epsilon_1$-close. At this point, $\mathbf{Z}'_h$ is a deterministic function of $\overline{\mathbf{Y}_1}'$, and hence we can fix it without affecting the distribution of $\mathbf{Z}_h$. This completes the proof. □

Claim 4.4.10. *Conditioned on $\overline{\mathbf{Z}}'$, the r.v $\overline{\mathbf{Z}}$ is $O(\epsilon_1 \log(n/\epsilon_1))$-close to uniform on average.*

Proof. Fix the r.v's $\mathbf{W}, \mathbf{W}', \overline{\mathbf{V}}_1, \overline{\mathbf{V}_1}', \mathbf{Y}_1, \mathbf{Y}'_1\overline{\mathbf{Y}_1}, \overline{\mathbf{Y}_1}', f(\mathbf{X})$. By Claim 4.4.9, the following hold:

- $\overline{\mathbf{Y}}$ has average conditional min-entropy at least $d' - 2d_3 - \log(1/\epsilon_1)$.
- for each $i \in [L]$, $\mathbf{Z}_i$ is $3\epsilon_1$-close to uniform on average.
- there exists $h \in [L]$ such that further conditioned on $\mathbf{Z}'_i$, the r.v $\mathbf{Z}_i$ is $3\epsilon_1$-close to uniform on average.
- $\{\overline{\mathbf{Y}}, \overline{\mathbf{Y}}'\}$ is independent of $\{\mathbf{Z}, \mathbf{Z}'\}$.

Let $d'' = 2d_3 + \log(1/\epsilon_1), r = \lceil \frac{\log L}{\log \ell} \rceil = \lceil \sqrt{\log L} \rceil$. Thus $d'' = O(\log(k/\epsilon_1)), r = O(\sqrt{\log\log(n/\epsilon_1)}), \ell = 2^{O(\sqrt{\log\log(n/\epsilon_1)})}$. In order to use Theorem 4.4.4, we observe that for a large enough constant $C_{4.4.3}$ the following hold:

- $\overline{\mathbf{Y}}$ has conditional min-entropy at least $d - d''$,
- $d' \geq (c_{3.5.7}\ell \log(m''/\epsilon_1) + d'')3^{r+1}$,
- $m < (0.9)^r(m'' - c_{3.5.7}\ell(t+1)r\log(m/\epsilon_1))$.

Thus the conditions of Theorem 4.4.4 are met, and hence it follows that conditioned on $\overline{\mathbf{Z}}'$, the r.v $\overline{\mathbf{Z}}$ is $2c'_{3.5.7}L\epsilon_1$-close to uniform on average. Recall that $L = O(\log(n/\epsilon_1))$, and hence the claim follows. □

□

4.4.3 A Trade-off Between Min-Entropy and Seed Length

We prove Theorem 5 and Theorem 6 in this section. Our main tool is a new NIPM construction which uses an even shorter seed but requires matrices with larger rows.

The main idea is to use our previous NIPM to construct a more involved NIPM, which can be used to give explicit non-malleable extractors with either a better seed length or a better min-entropy requirement. For simplicity and clarity, we will just assume $t = 1$, i.e., there is only one tampering adversary. This is also the most interesting case for standard privacy amplification protocols.

Note that our previous NIPM construction implies the following theorem.

Theorem 9. *For all integers $m, L > 0$, any $\epsilon > 0$, there exists an explicit $(L, 1, 0, \epsilon, \epsilon')$-NIPM : $\{0,1\}^{mL} \times \{0,1\}^d \to \{0,1\}^{m'}$, where $d = 2^{O(\sqrt{\log L})}\log(m/\epsilon), m' = \frac{m}{2^{\sqrt{\log L}}} - 2^{O(\sqrt{\log L})}\log(m/\epsilon)$ and $\epsilon' = O(\epsilon L)$.*

We start by proving the following lemma.

Lemma 4.4.11. *For all integers $m, L > 0$, any $\epsilon > 0$, if there is an explicit $(L, 1, 0, \epsilon, \epsilon_1)$-NIPM$_1$: $\{0,1\}^{mL} \times \{0,1\}^{d_1} \to \{0,1\}^{m_1}$, with $d_1 \leq 2^{r(\log L)^{1/q}}\log(m/\epsilon), m_1 = \frac{m}{2^{s(\log L)^{1-1/q}}} - 2^{O((\log L)^{1-1/q})}\log(m/\epsilon)$ and $\epsilon_1 \leq g(L)\epsilon L$, where $g(L)$ is a monotonic non-decreasing function*

of L, and r, s, q are parameters, with $q \in \mathbb{N}$, then there is an explicit $(L, 1, 0, \epsilon, \varepsilon_2)$-NIPM$_2$: $\{0,1\}^{mL} \times \{0,1\}^{d_2} \to \{0,1\}^{m_2}$, with $d_2 \leq 2^{2r^{1-1/(q+1)}(\log L)^{\frac{1}{q+1}}} \log(m/\epsilon), m_2 = \frac{m}{2^{sr^{\frac{1}{q+1}}(\log L)^{1-1/(q+1)}}} - 2^{O((\log L)^{1-1/(q+1)})} \log(m/\epsilon))$ and $\varepsilon_2 \leq 2\varepsilon_1$.

Proof. The idea is to use Algorithm 6, with $(\ell, 1, 0, \epsilon, \epsilon_1')$-NIPM$_1$, $\epsilon_1' \leq g(\ell)\ell\epsilon$ as the simpler merger for some parameter ℓ s.t. in each step, the merger acts on ℓ rows. Following the proof of Theorem 4.4.4, it can be shown that the seed length of NIPM$_2$ will be

$$d_2 = \log(m/\epsilon) 2^{r(\log \ell)^{1/q}} 2^{2\frac{\log L}{\log \ell}}.$$

We now choose an ℓ to minimize this, which gives $(\log \ell)^{\frac{q+1}{q}} = \frac{2 \log L}{r}$, and thus the seed length is

$$d_2 = 2^{2r^{\frac{q}{q+1}}(2\log L)^{\frac{1}{q+1}}} \log(m/\epsilon).$$

It can be verified that for this setting of parameters, the output length is

$$\begin{aligned} m_2 &= \frac{m}{\left(2^{s(\log L)^{1-1/q}}\right)^{\frac{\log L}{\log \ell}}} - O(\ell \log(m/\epsilon)) \\ &= \frac{m}{2^{sr^{\frac{1}{q+1}}(\log L)^{\frac{q}{q+1}}}} - 2^{O((\frac{\log L}{r})^{\frac{q}{q+1}})} \log(m/\epsilon) \\ &= \frac{m}{2^{sr^{\frac{1}{q+1}}(\log L)^{1-1/(q+1)}}} - 2^{O((\log L)^{\frac{q}{q+1}})} \log(m/\epsilon)) \end{aligned}$$

Finally, the error is bounded by $\sum_{i=1}^{\frac{\log L}{\log \ell}} g(\ell)\epsilon\ell^i < 2g(\ell)L\epsilon < 2\epsilon_1$. □

Now, starting with the NIPM from Theorem 9, and using Lemma 4.4.11 an optimal number of times, we have the following theorem.

Theorem 10. *For all integers $m, L > 0$, any $\epsilon > 0$, there exists an explicit $(L, 1, 0, \epsilon, \epsilon')$-NIPM : $\{0,1\}^{mL} \times \{0,1\}^{d} \to \{0,1\}^{m'}$, where $d = 2^{O(\sqrt{\log \log L})} \log(m/\varepsilon), m' = \frac{m}{L2^{(\log \log L)^{O(1)}}} - O(L \log(m/\varepsilon))$ and $\epsilon' = 2^{O(\sqrt{\log \log L})} L\varepsilon$.*

Proof. We start from the basic case with the $(L, 1, 0, \epsilon, \epsilon')-$NIPM from Theorem 9. Thus $q = 2, r = O(1), s = 1$. We now use Lemma 4.4.11, increasing q by one each time. Eventually, we stop at $q = \sqrt{\log \log L}$, noticing that this minimize the seed length. It can be verified that the seed length of the final NIPM is $2^{O(\sqrt{\log \log L})} \log(m/\varepsilon)$, the output length is $\frac{m}{L2^{(\log \log L)^{O(1)}}} - O(L \log(m/\varepsilon))$ and the error is bounded by $\varepsilon \leq 2^{O(\sqrt{\log \log L})} L\varepsilon$. □

Using the NIPM from Theorem 10 in Algorithm 7, we obtain the following non-malleable extractor with a slightly shorter seed length than Theorem 4.4.3 at the expense of requiring larger min-entropy.

Theorem 6 (restated)**.** *For all $n, k \in \mathbb{N}$ and any $\epsilon > 0$, with $k \geq (\log(n/\epsilon))^3 2^{(\log \log \log(n/\epsilon))^{O(1)}}$, there exists an explicit (k, ϵ)-non-malleable extractor* $\mathrm{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$*, where* $d = \log(n/\epsilon) 2^{2^{O(\sqrt{\log \log \log(n/\epsilon)})}}, m = \frac{k}{\log(n/\epsilon) 2^{(\log \log \log(n/\epsilon))^{O(1)}}} - O((\log(n/\epsilon))^2)$.

The proof of Theorem 6 is exactly similar to Theorem 4.4.3, and we skip it.

It is not hard to modify Algorithm 7 such that the the role of the source and the seed are swapped, in the sense that the seed to NIPM is a deterministic function of the source to the non-malleable extractor, and the matrix is a deterministic function of the seed to the non-malleable extractor. By this modification. we can achieve a non-malleable extractor that works for lower slightly min-entropy than Theorem 4.4.3 at the expense of using a larger seed.

4.5 Improved t-Non-Malleable Extractors

The framework to construct non-malleable extractors in Section 4.4 can be generalized directly to construct non-malleable extractors that can handle multiple adversaries.

In particular, Theorem 4.4.6 generalizes to the case there are t tampered variables, and further our NIPM construction in Theorem 3.5.7 handles t adversaries. By using these versions of the components in the above construction, the following theorem is easy to obtain. Since the proof is similar to the proof of Theorem 4.4.3, we omit the proof of the following theorem.

Theorem 11. *There exists a constant $\delta > 0$ such that for all $n, k, t, \ell \in \mathbb{N}$ and any $\epsilon > 0$, with $r = (\log\log(n/\epsilon))/(\log \ell)$, $k = \Omega(t^{2r}\ell \log(n/\epsilon))$, there exists an explicit (t, k, ϵ)-non-malleable extractor* $\mathrm{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$*, where $d = O(t^{(1+\delta)r}\ell \log(n/\epsilon))$ and $m = (\delta k - \ell t r \log(n/\epsilon))/(2t)^{(\log L/\log \ell)}$.*

Chapter 5

Resilient Functions and Extracting from NOBF Sources

[1] Ben-Or and Linial [BL85] first studied resilient functions when they introduced the perfect information model. In the simplest version of this model, there are n computationally unbounded players that can each broadcast a bit once. At the end, some function is applied to the broadcast bits. In the collective coin-flipping problem, the output of this function should be a nearly-random bit. The catch is that some malicious coalition of players may wait to see what the honest players broadcast before broadcasting their own bits. Thus, a resilient function is one where the bit is unbiased even if the malicious coalition is relatively large (but not too large).

This model can be generalized to allow many rounds, and has been well studied [BL85, KKL88, Sak89, AL93, AN93, BN96, RZ01, Fei99, RSZ02]; also see the survey by Dodis [Dod06]. Resilient functions correspond to 1-round protocols.

To formally define resilient functions, we introduce the notion of influence of sets on functions.

Definition 5.0.1 (Influence of a set)**.** *Let $f : \{0,1\}^n \to \{0,1\}$ be any Boolean function on variables*

[1] parts of this chapter have been previously published [CZ16a]

$x_1, \ldots, x_n$. *The influence of a set* $Q \subseteq \{x_1, \ldots, x_n\}$ *on* f*, denoted by* $\mathbf{I}_Q(f)$*, is defined to be the probability that* f *is undetermined after fixing the variables outside* Q *uniformly at random. Further, for any integer* q *define* $\mathbf{I}_q(f) = \max_{Q \subseteq \{x_1,\ldots,x_n\}, |Q|=q} \mathbf{I}_Q(f)$*. More generally, let* $\mathbf{I}_{Q,\mathcal{D}}(f)$ *denote the probability that* f *is undetermined when the variables outside* Q *are fixed by sampling from the distribution* $\mathcal{D}$*. We define* $\mathbf{I}_{Q,t}(f) = \max_{\mathcal{D} \in \mathcal{D}_t} \mathbf{I}_{Q,\mathcal{D}}(f)$*, where* $\mathcal{D}_t$ *is the set of all* t*-wise independent distributions. Similarly,* $\mathbf{I}_{Q,t,\gamma}(f) = \max_{\mathcal{D} \in \mathcal{D}_{t,\gamma}} \mathbf{I}_{Q,\mathcal{D}}(f)$ *where* $\mathcal{D}_{t,\gamma}$ *is the set of all* (t,γ)*-wise independent distributions. Finally, for any integer* q *define* $\mathbf{I}_{q,t}(f) = \max_{Q \subseteq \{x_1,\ldots,x_n\}, |Q|=q} \mathbf{I}_{Q,t}(f)$ *and* $\mathbf{I}_{q,t,\gamma}(f) = \max_{Q \subseteq \{x_1,\ldots,x_n\}, |Q|=q} \mathbf{I}_{Q,t,\gamma}(f)$.

Definition 5.0.2 (Resilient Function)**.** *Let* $f : \{0,1\}^n \to \{0,1\}$ *be any Boolean function on variables* $x_1, \ldots, x_n$ *and* q *any integer. We say* f *is* (q,ϵ)*-resilient if* $\mathbf{I}_q(f) \leq \epsilon$*. More generally, we say* f *is* t*-independent* (q,ϵ)*-resilient if* $\mathbf{I}_{q,t}(f) \leq \epsilon$ *and* f *is* (t,γ)*-independent* (q,ϵ)*-resilient if* $\mathbf{I}_{q,t,\gamma}(f) \leq \epsilon$.

Resilient functions have applications in extracting from bit-fixing sources. Roughly, a bit-fixing source is a source where some subset of the bits are fixed and the remaining ones chosen in some random way. Usually these remaining bits are chosen uniformly at random, but in this chapter we also consider the case when they are chosen t-wise independently. Extraction is easier if the fixed bits cannot depend on the random bits. Such sources are called oblivious bit-fixing sources, and have been investigated in a line of work [CGH$^+$85, KZ07a, GRS06, Rao09b]. The best known explicit extractors for oblivious sources work for min-entropy at least $\log^C(n)$ with exponentially small error [Rao09b], and from arbitrary min-entropy with polynomially small error [KZ07a]. They have applications to cryptography [CGH$^+$85, KZ07a].

Resilient functions immediately give an extractor for the more difficult family of non-oblivious bit-fixing sources, where the fixed bits may depend on the random bits. We formally record this connection.

Definition 5.0.3 (Non-Oblivious Bit-Fixing Sources)**.** *A source* $\mathbf{Z}$ *on* $\{0,1\}^n$ *is called a* (q,t,γ)*-non-oblivious bit-fixing source (NOBF source for short) if there exists a subset of coordinates* $Q \subseteq [n]$

of size at most q such that the joint distribution of the bits indexed by $\overline{Q} = [n] \setminus Q$ is (t, γ)-wise independent. The bits in the coordinates indexed by Q are allowed to depend arbitrarily on the bits in the coordinates indexed by $\overline{Q}$.

Lemma 5.0.4. *Let $f : \{0,1\}^n \to \{0,1\}$ be a Boolean function such that for any t-wise independent distribution $\mathcal{D}$, $|\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[f(x)] - \frac{1}{2}| \leq \epsilon_1$. Suppose for some $q > 0$, $\mathbf{I}_{q,t}(f) \leq \epsilon_2$. Then, f is an extractor for (q, t, γ)-NOBF sources on n bits with error $\epsilon_1 + \epsilon_2 + \gamma n^t$.*

The results in this chapter are based on joint work with David Zuckerman [CZ16a].

5.1 Our Results and Overview of Techniques

As discussed above, since resilient functions have applications in distributed computing and also extracting from NOBF sources, it is important to have explicit constructions of such functions. For $t < \sqrt{n}$, the only known function that is t-independent (q, ϵ_1)-resilient function is the majority function [DGJ+10, Vio14] for $t = O(1)$ and $q < n^{\frac{1}{2}-\tau}$, $\tau > 0$.

However, for larger t, there are better known resilient functions. In particular, the iterated majority function of Ben-Or and Linial handles a larger $q = O(n^{\log_3 2})$ for $t = n$, but it is not clear if it remains resilient for smaller t. Further, Ajtai and Linial [AL93] showed the existence of functions that are resilient for $q = O(n/\log^2 n)$ and $t = n$. However, their functions are not explicit and require time $n^{O(n^2)}$ to deterministically construct. We note here that by a result in [KKL88], the largest q one can hope to handle is $O(n/\log n)$.

Our main result on resilient functions is the following.

Theorem 12. *There exists a constant c such that for any $\delta > 0$ and every large enough integer $n \in \mathbb{N}$, there exists an efficiently computable monotone Boolean function $f : \{0,1\}^n \to \{0,1\}$ satisfying: For any $q > 0, t \geq c(\log n)^{18}$,*

- *f is a depth 4 circuit of size $n^{O(1)}$.*
- *For any (t, γ)-wise independent distribution $\mathcal{D}$, $|\mathbf{E}_{\mathbf{x}\sim\mathcal{D}}[f(\mathbf{x})] - \frac{1}{2}| \leq \frac{1}{n^{\Omega(1)}}$.*

- $\mathbf{I}_{q,t}(f) \leq q/n^{1-\delta}$.

Our main result on extracting from bit-fixing sources is the following. We note that this direct from Theorem 12 and Lemma 5.0.4.

Theorem 13. *There exists a constant c such that for any constant $\delta > 0$, and for all $n \in \mathbb{N}$, there exists an explicit extractor* $\text{bitExt} : \{0,1\}^n \to \{0,1\}$ *for the class of (q,t,γ)-non-oblivious bit-fixing sources with error $n^{-\Omega(1)}$, where $q \leq n^{1-\delta}, t \geq c\log^{18}(n)$ and $\gamma \leq 1/n^{t+1}$.*

Subsequent Work: Meka [Mek15] built on our ideas to construct a resilient function matching the probabilistic construction of Ajtai-Linial.

We now outline the main ideas in the proof of Theorem 12. We first show that if the function f is monotone, in AC^0 and almost unbiased, then it is enough to bound $\mathbf{I}_q(f)$ to show that f satisfies the conclusions of Theorem 12. The key observation is the following simple fact: for any set of variables Q, it is possible to check using another small AC^0 circuit $\mathcal{E}$ if the function f is undetermined for some setting of the variables outside Q. This crucially relies on the fact that f is monotone. Next, using the result of Braverman [Bra10] that bounded independence fools small AC^0 circuits, we conclude that the bias of the circuit $\mathcal{E}$ is roughly the same when the variables outside Q are drawn from a bounded-independence distribution, and when they are drawn from the uniform distribution. The conclusion now follows using the bound on $\mathbf{I}_Q(f)$.

Thus all that remains is to construct a small monotone AC^0 circuit f, that is almost balanced under the uniform distribution, and $\mathbf{I}_q(f) = o(1)$ for $q < D^{1-\delta}$. The high level idea for this construction is to derandomize the probabilistic construction of Ajtai-Linial using extractors. The tribes function introduced by Ben-Or and Linial [BL85] is a disjunction taken over AND's of equi-sized blocks of variables. The Ajtai-Linial function is essentially a conjunction of non-monotone tribes functions, with each tribes function using a different partition and the variables in each tribes function being randomly negated with probability $1/2$, and the partitions are chosen according to the probabilistic method. We sketch informally our ideas to derandomize this construction. For each $i \in [R]$, let P^i be a equi-partition of $[n]$, $n = MB$, into blocks of size B. Let P^i_j denote the

j'th block in P^i. Define,

$$f(x) = \bigwedge_{1 \le i \le R} \bigvee_{1 \le j \le M} \bigwedge_{\ell \in P^i_j} x_\ell.$$

First, we abstract out properties that these partitions need to satisfy for f to be almost unbiased and also $(n^{1-\delta}, \epsilon)$-resilient. Informally, we show that

1. If for all i_1, i_2, j_1, j_2 with $(i_1, j_1) \neq (i_2, j_2)$, $|P^{i_1}_{j_1} \cap P^{i_2}_{j_2}| \le 0.9B$, then f is almost unbiased,
2. If for any set Q of size $q < n^{1-\delta}$, the number of partitions P^i containing a block P^i_j such that $|P^i_j \cap Q| > \delta B/2$ is $o(R)$, then f is $(n^{1-\delta}, \epsilon)$-resilient.

An ingredient in the proof of (1) is Janson's inequality (see Theorem 5.3.22).

It is important that unlike in Ajtai-Linial and earlier modifications [RZ01], we don't need to negate variables, and thus f is monotone.

The second property seems related to the property of extractors captured in Theorem 2.4.1. However, it is not obvious how to use such extractors to construct these partitions. We construct a family of equi-partitions from a seeded extractor $\mathrm{Ext} : \{0,1\}^r \times \{0,1\}^b \to \{0,1\}^m$ as follows. Each P^w corresponds to some $w \in \{0,1\}^r$. One block of P^w is $P^w_{\vec{0}} = \{(y, \mathrm{Ext}(x,y)) : y \in \{0,1\}^b\}$. The other block are shifts of this, i.e., for any $s \in \{0,1\}^m$, define $P^w_s = \{(y, \mathrm{Ext}(x,y) \oplus s) : y \in \{0,1\}^b\}$. This gives $R = 2^r$ partitions of $[n]$ with $n = 2^{m+b}$.

For any good enough extractor, we show that (2) is satisfied using a basic property of extractors and an averaging argument. To show that the partitions satisfy (1), we need an additional property of the extractor, which informally requires us to prove that the intersection of any two arbitrary shifts of neighbors of any two distinct nodes $w_1, w_2 \in \{0,1\}^r$ in G_{Ext} is bounded. This essentially is a strong variant of a design extractor of Li [Li12a]. We show that Trevisan's extractor has this property. This completes the informal sketch of our resilient function construction. We note that our actual construction is slightly more complicated and is a depth 4 circuit. The extra layer enables us to simulate each of the bits $x_1, \ldots, x_n$ having $\Pr[x_1 = 1]$ close to 1, which we need to make f almost unbiased.

5.2 Monotone Constant-Depth Resilient Functions are t-Independent Resilient

We show if f is a constant depth monotone circuit, then in order to prove an upper bound for $\mathbf{I}_{q,t,\gamma}(f)$, it is in fact enough to upper bound $\mathbf{I}_q(f)$, which is a simpler quantity to handle.

Theorem 5.2.1. *There exists a constant $b > 0$ such that the following holds: Let $\mathcal{C} : \{0,1\}^n \to \{0,1\}$ be a monotone circuit in AC^0 of depth d and size m such that $|\mathbf{E}_{x\sim\mathbf{U}_n}[\mathcal{C}(x)]-\frac{1}{2}| \le \epsilon_1$. Suppose $q > 0$ is such that $\mathbf{I}_q(\mathcal{C}) \le \epsilon_2$. If $t \ge b(\log(5m/\epsilon_3))^{3d+6}$, then $\mathbf{I}_{q,t}(\mathcal{C}) \le \epsilon_2+\epsilon_3$ and $\mathbf{I}_{q,t,\gamma}(\mathcal{C}) \le \epsilon_2+\epsilon_3+\gamma n^t$. Further, for any distribution $\mathcal{D}$ that is (t,γ)-wise independent, $|\mathbf{E}_{x\sim\mathcal{D}}[\mathcal{C}(x)]-\frac{1}{2}| \le \epsilon_1+\epsilon_3+\gamma n^t$.*

An important ingredient in the our proof is a result Braverman [Bra10], which was recently refined by Tal [Tal14].

Theorem 5.2.2 ([Bra10] [Tal14])**.** *Let $\mathcal{D}$ be any $t = t(m,d,\epsilon)$-wise independent distribution on $\{0,1\}^n$. Then for any circuit $\mathcal{C} \in AC^0$ of depth d and size m,*

$$|\mathbf{E}_{x\sim\mathbf{U}_n}[\mathcal{C}(x)] - \mathbf{E}_{x\sim\mathcal{D}}[\mathcal{C}(x)]| \le \epsilon$$

where $t(m,d,\epsilon) = O(\log(m/\epsilon))^{3d+3}$.

We also recall a result about almost t-wise independent distributions.

Theorem 5.2.3 ([AGM03])**.** *Let $\mathcal{D}$ be a (t,γ)-wise independent distribution on $\{0,1\}^n$. Then there exists a t-wise independent distribution that is $n^t\gamma$-close to $\mathcal{D}$.*

Proof of Theorem 5.2.1. The bound on $\mathbf{E}_{x\sim\mathcal{D}}[\mathcal{C}(x)]$ is direct from Theorem 5.2.2 and Theorem 7.3.4. We now proceed to prove the influence property.

Consider any set Q of variables, $|Q| = q$. Let $\overline{Q} = [n] \setminus Q$. We construct a function $\mathcal{E}_Q : \{0,1\}^{n-q} \to \{0,1\}$ such that $\mathcal{E}_Q(y) = 1$ if and only if $\mathcal{C}$ is undetermined when $x_{\overline{Q}}$ is set to y.

Thus, it follows that

$$\mathbf{E}_{y\sim\mathbf{U}_{n-q}}[\mathcal{E}_Q(y)] = \mathbf{Pr}_{y\sim\mathbf{U}_{n-q}}[\mathcal{E}_Q(y)=1] = \mathbf{I}_Q(\mathcal{C}) \leq \epsilon_2.$$

Let $\mathcal{D}$ be any t-wise independent distribution. We have,

$$\mathbf{E}_{y\sim\mathcal{D}}[\mathcal{E}_Q(y)] = \mathbf{Pr}_{y\sim\mathcal{D}}[\mathcal{E}_Q(y)=1] = \mathbf{I}_{Q,\mathcal{D}}(\mathcal{C}).$$

Thus to prove that $\mathbf{I}_{Q,\mathcal{D}}(\mathcal{C}) \leq \epsilon_2 + \epsilon_3$, it is enough to prove that

$$|\mathbf{E}_{y\sim\mathbf{U}_{n-q}}[\mathcal{E}_Q(y)] - \mathbf{E}_{y\sim\mathcal{D}}[\mathcal{E}_Q(y)]| \leq \epsilon_3. \tag{5.1}$$

We construct $\mathcal{E}_Q$ as follows: Let $\mathcal{C}_0$ be the circuit obtained from $\mathcal{C}$ by setting all variables in Q to 0. Let $\mathcal{C}_1$ be the circuit obtained from $\mathcal{C}$ by setting all variables in Q to 1. Define $\mathcal{E}_Q \coloneqq \neg(\mathcal{C}_0 = C_1)$. Since $\mathcal{C}$ is monotone, $\mathcal{E}_Q$ satisfies the required property. Further $\mathcal{E}_Q$ can be computed by a circuit in AC^0 of depth $d+2$ and size $4m+3$. It can be checked that the depth of $\mathcal{E}_Q$ can be reduced to $d+1$ by combining two layers. Thus (5.1) now directly follows from Theorem 5.2.2. The bound on $\mathbf{I}_{\mathcal{C},t,\gamma}(q)$ follows from an application of Theorem 7.3.4. □

5.3 Monotone Boolean Functions in AC^0 Resilient to Coalitions

The main result in this section is an explicit construction of a constant depth monotone circuit f which is resilient to coalitions and is almost balanced under the uniform distribution. This is the final ingredient in our construction of a 2-source extractor.

Theorem 5.3.1. *For any $\delta > 0$, and every large enough integer n, there exists a polynomial time computable monotone Boolean function $f : \{0,1\}^n \to \{0,1\}$ satisfying:*

- *f is a depth 4 circuit in AC^0 of size $n^{O(1)}$.*
- *$\left|\mathbf{E}_{x\sim\mathbf{U}_n}[f(x)] - \frac{1}{2}\right| \leq \frac{1}{n^{\Omega(1)}}$.*

- *For any* $q > 0$, $\mathbf{I}_q(f) \leq q/n^{1-\delta}$.

We first prove Theorem 12, which follows easily from the above theorem.

Proof of Theorem 12. Let $f : \{0,1\}^n \to \{0,1\}$ be the function from Theorem 5.3.1 such that for any $q > 0$, $\mathbf{I}_q(f) \leq q/n^{1-\frac{\delta}{2}}$. Also we have that f is monotone and is a depth 4 AC^0 circuit.

Fix $\epsilon_3 = 1/n$. Thus by Theorem 5.2.1, it follows that there exists a constant b such that for any $t \geq b(\log(5n/\epsilon_3))^{18}$, $q > 0$,

$$\mathbf{I}_{q,t,\gamma}(f) \leq \epsilon_3 + \frac{q}{n^{1-\frac{\delta}{2}}} \leq \frac{q}{n^{1-\delta}}.$$

Further, using Theorem 5.2.1, for any t-wise independent distribution $\mathcal{D}$, we have

$$\left|\mathbf{E}_{x\sim\mathcal{D}}[f(x)] - \frac{1}{2}\right| \leq \frac{1}{n} + \frac{1}{n^{\Omega(1)}}.$$

□

The remainder of this section is used to prove Theorem 5.3.1. Our starting point is the work of Ajtai and Linial [AL93], who proved the existence of functions computable by linear sized depth 3 circuits in AC^0 that are $(\Omega(n/\log^2 n), \epsilon)$-resilient. However, this construction is probabilistic, and deterministically finding such functions requires time $n^{O(n^2)}$. Further these functions are not guaranteed to be monotone (or even unate). We provide some intuition of our construction in the introduction.

We initially construct a depth 3 circuit which works, but then the inputs have to be chosen from independent Bernoulli distributions where the probability p of 1 is very different from 1/2. By observing that we can approximate this Bernoulli distribution with a CNF on uniform bits, we obtain a depth 4 circuit which works for uniformly random inputs.

5.3.1 Our Construction and Key Lemmas

Construction 1: Let $\text{Ext} : \{0,1\}^r \times \{0,1\}^b \to \{0,1\}^m$ be a strong-seeded extractor set to extract from min-entropy $k = 2\delta r$ with error $\epsilon \leq \delta/4, b = \delta_1 m$, $\delta_1 = \delta/20$, and output length $m = \delta r$. Assume that Ext is such that $\epsilon > 1/M^{\delta_1}$. Let $R = 2^r$, $B = 2^b$, $M = 2^m$ and $K = 2^k$. Let $s = BM$. Thus $s = M^{1+\delta_1}$.

Let $\{0,1\}^r = \{v_1, \ldots, v_R\}$. We define a collection of R equi-partitions of $[s]$, $\mathcal{P} = \{P^{v_1}, \ldots, P^{v_R}\}$ as follows: Let G_{Ext} be the bipartite graph corresponding to Ext and let $\mathcal{N}(x)$, for any $x \in \{0,1\}^r$, denote the neighbours of x in G_{Ext}. For some $v \in \{0,1\}^r$, let $\mathcal{N}(v) = \{z_1, \ldots, z_B\}$. For each $w \in \{0,1\}^m$, the set $\{(j, z_j \oplus w) : j \in \{0,1\}^b\}$ is defined to be a block in P^v, where $\oplus$ denotes the bit-wise XOR of the two strings. Note that P^v indeed forms an equi-partition of $[s]$ with M blocks of size B.

Define the function $f_{\text{Ext}} : \{0,1\}^s \to \{0,1\}$ as:

$$f_{\text{Ext}}(y) = \bigwedge_{1 \leq i \leq R} \bigvee_{1 \leq j \leq M} \bigwedge_{\ell \in P^i_j} y_\ell.$$

Let

$$\gamma = \frac{\ln M - \ln \ln(R/\ln 2)}{B}.$$

We prove the following lemmas from which the proof of Theorem 5.3.1 is straightforward. We first introduce some definitions.

Definition 5.3.2 ((n,τ)-Bernoulli distribution)**.** *A distribution on n bits is an (n,τ)-Bernoulli distribution, denoted by $\mathbf{Ber}(n,\tau)$, if each bit is independently set to 1 with probability τ and set to 0 with probability $1-\tau$.*

Lemma 5.3.3. *Let $\text{Ext} : \{0,1\}^r \times \{0,1\}^b \to \{0,1\}^m$ be the extractor used in Construction 1. For any constant $\epsilon_1 > 0$, let $(1 - B^{-\epsilon_1})\gamma \leq p_1 \leq \gamma$. Then there exists a constant $\delta > 0$ such that for any $q > 0$,*

$$\mathbf{I}_{q,\mathbf{Ber}(s,1-p_1)}(f_{\text{Ext}}) \leq \frac{q}{s^{1-\delta}}.$$

The following generalizes the notion of a design extractor which was introduced by Li [Li12a].

Definition 5.3.4 (Shift-design extractor)**.** *Let* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a strong-seeded extractor. Let* $n = 2^d$*. If for any distinct* $x, x' \in \{0,1\}^n$*, and arbitrary* $y, y' \in \{0,1\}^m$

$$|\{(h, \mathrm{Ext}(x,h) \oplus y) : h \in \{0,1\}^d\} \cap \{(h, \mathrm{Ext}(x',h) \oplus y') : h \in \{0,1\}^d\}| \leq (1-\eta)n,$$

then Ext *is called an* η*-shift-design extractor.*

Lemma 5.3.5. *Let* $\mathrm{Ext} : \{0,1\}^r \times \{0,1\}^b \to \{0,1\}^m$ *be the extractor used in Construction* 1*. Suppose* Ext *is a* $\frac{1}{10}$*-shift-design extractor. For any constant* $\epsilon_1 > 0$*, let* $(1 - B^{-\epsilon_1})\gamma \leq p_1 \leq \gamma$*. Then, the following holds:*

$$\left| \mathbf{E}_{y \sim \mathbf{Ber}(s, 1-p_1)}[f_{\mathrm{Ext}}(y)] - \frac{1}{2} \right| \leq B^{-\Omega(1)}.$$

Lemma 5.3.6. *Let* $\mathrm{TExt} : \{0,1\}^r \times \{0,1\}^b \to \{0,1\}^m$ *be the Trevisan extractor from Theorem* 2.1.5 *with parameters as in Construction* 1*. Then,* TExt *is a* $\frac{1}{10}$*-shift-design extractor.*

Lemma 5.3.7. *Suppose* $\gamma < 9/10$*. Then for any* $\nu > 0$*, there exists an explicit size* h *monotone* CNF $\mathcal{C}$ *on* h *bits, where* $h = O\left(\frac{1}{\nu} \ln\left(\frac{1}{\nu}\right)\right)$*, such that* $\gamma - \nu \leq \mathbf{Pr}_{x \sim \mathbf{U}_h}[\mathcal{C}(x) = 0] < \gamma$*.*

We first show how to derive Theorem 5.3.1 from the above lemmas.

Proof of Theorem 5.3.1. Let $\mathrm{TExt} : \{0,1\}^r \times \{0,1\}^b \to \{0,1\}^m$ be the Trevisan extractor from Theorem 2.1.5 with parameters as in Construction 1: $k = 2\delta r, m = \delta r, \delta_1 = \delta/20$ and $\epsilon = 2^{-\delta_2\sqrt{r}}$ where δ_2 is chosen appropriately such that the seed length of TExt from Theorem 2.1.5 is (for some constant λ)

$$b = \frac{\lambda \log^2(r/\epsilon)}{\log(k/m)} = \frac{\lambda \log^2(r/2^{-\delta_2\sqrt{r}})}{\log 2} = \lambda(\delta_2^2 r + \log^2 r + 2\delta_2\sqrt{r}\log r) = \delta_1\delta r = \delta_1 m.$$

Thus, indeed $M^{-\delta_1} < \epsilon < \delta/4$.

We now fix the parameter r as follows. Let the parameter ν in Lemma 5.3.7 be set to γ/B^{ϵ_1}, where $\epsilon_1 = \delta/4$ and let $\mathcal{C}$ be the size h monotone CNF circuit guaranteed by Lemma 5.3.7, where $h < B^{1+2\epsilon_1}$. Thus, $(1 - B^{-\epsilon_1})\gamma \leq \mathbf{Pr}_{x\sim\mathbf{U}_h}[\mathcal{C}(x) = 0] < \gamma$.

Choose the largest integer r such that for $m = \delta r$, we have $n' = sh = BMh < n$. It follows that for this choice of r, $n' = \Omega(n)$. We construct our function on n' bits. The size of the coalition is at most $n^{1-\delta} = (n')^{1-\delta'}$, where $\delta' = \delta - o(1)$. Thus, we may assume $n = n' = BMh$ and $\delta = \delta'$. Thus $n = BMh < M^{1+\delta_1+(1+2\epsilon_1)\delta_1}$ and $B = n^{\Omega(1)}$.

We now use Construction 1 and construct the function $f_{\mathrm{TExt}} : \{0,1\}^s \to \{0,1\}$, where we instantiate Ext with extractor TExt as set up above. Let f be the function derived from f_{TExt} by replacing each variable y_i by a copy of the monotone CNF $\mathcal{C}$ set up above. Since TExt is a polynomial time function, f_{TExt} can be constructed in polynomial time. Thus f is computable by a polynomial time algorithm. Further, f is an $O(RMBh) = n^{O(1)}$ sized monotone circuit in AC^0 of depth 4.

We observe that,

$$\begin{aligned} s^{1-\frac{\delta}{2}} &= (MB)^{1-\frac{\delta}{2}} \\ &> (MB)^{(1+\frac{\delta}{2})(1-\delta)} \\ &> (MB^3)^{1-\delta} && \text{(since } M^{\delta/2} > B^2\text{)} \\ &\geq (MBh)^{1-\delta} = n^{1-\delta}. \end{aligned}$$

This calculation and Lemma 5.3.7 yields that

$$\mathbf{I}_{n^{1-\delta}}(f) \leq \mathbf{I}_{s^{1-\frac{\delta}{2}},\mathbf{Ber}(s,1-p_1)}(f_{\mathrm{TExt}}).$$

Using Lemma 5.3.3, it follows that

$$\mathbf{I}_{q,\mathbf{Ber}(s,1-p_1)}(f_{\mathrm{Ext}}) \leq \frac{q}{s^{1-\frac{\delta}{2}}} < \frac{q}{n^{1-\delta}}.$$

We now bound the bias of f. By Lemma 5.3.6, we have that TExt is a $\frac{1}{10}$-shift-design extractor. Thus by Lemma 5.3.5, we have

$$\left|\mathbf{E}_{y\sim\mathbf{Ber}(s,1-p_1)}[f_{\mathrm{TExt}}(y)] - \frac{1}{2}\right| \leq B^{-\Omega(1)} = n^{-\Omega(1)}.$$

Finally, using Lemma 5.3.7, it follows that

$$\left|\mathbf{E}_{x\sim\mathbf{U}_n}[f(x)] - \frac{1}{2}\right| \leq \frac{1}{n^{\Omega(1)}}$$

. □

Proof of Lemma 5.3.6. To prove that TExt is a $\frac{1}{10}$-shift-design extractor, we first recall the construction of the Trevisan extractor $\mathrm{TExt} : \{0,1\}^r \times \{0,1\}^b \to \{0,1\}^m$.

For any input $y \in \{0,1\}^r$, we describe the construction of the Trevisan extractor [Tre01, RRV02] to obtain the first bit of the output since this is enough for the purpose of this proof. Fix an asymptotically good binary linear error correcting code $\mathcal{C}'$ with constant relative rate α, block length $\bar{r} = (r+1)/\alpha$, and relative distance $\frac{1}{2} - \beta$, where $\beta < \epsilon$. Further assume that $\mathcal{C}'$ contains the all 1's string $\vec{1}$. Let $\{v_1, \ldots, v_{r+1}\}$ be a basis of C' with $v_{r+1} = \vec{1}$. Let $\mathcal{C}$ be the binary linear code generated by $\{v_1, \ldots, v_r\}$ i.e., $\mathcal{C} = \mathrm{span}\{v_1, \ldots, v_r\}$. It follows that $\mathcal{C}$ does not contain $\vec{1}$, has relative rate $\alpha(1 - \frac{1}{\bar{r}}) > 0.9\alpha$ and relative distance $\frac{1}{2} - \beta$. Let $\mathrm{Enc} : \{0,1\}^r \to \{0,1\}^{\bar{r}}$ be the encoding function of $\mathcal{C}$.

Further fix a subset $S_1 \subset [b]$ of size $\log(\bar{r})$. Then the first bit of the output of TExt on input y and seed z is the bit at the z_{S_1}'th coordinate of the string $c_y = \mathrm{Enc}(y)$. Thus, as we cycle over all seeds z, each bit of the string c_y appears equally often.

For any $x \in \{0,1\}^r$, define

$$T_x^0 = \{(h, \mathrm{TExt}(x,h)_{[1]}) : h \in \{0,1\}^b\}, \qquad T_x^1 = \{(h, \mathrm{TExt}(x,h)_{[1]} \oplus 1) : h \in \{0,1\}^b\}.$$

Let x, x' be any two distinct r bit strings. It follows by our argument above, and the fact that C'

is a linear code with distance $\frac{1}{2} - \beta$ containing $\vec{1}$ that $|T_x^{b_1} \cap T_{x'}^{b_2}| \leq (\frac{1}{2} + \beta)B < 0.9B$ for any two bits b_1 and b_2.

Let $y, y' \in \{0,1\}^m$. Let the first bit of y be b_1 and the first bit of y' be b_2. Thus,

$$|\{(h, \mathrm{TExt}(x,h) \oplus y) : h \in \{0,1\}^b\} \cap \{(h, \mathrm{TExt}(x',h) \oplus y') : h \in \{0,1\}^b\}| \leq |T_x^{b_1} \cap T_{x'}^{b_2}| \leq 0.9B.$$

□

Proof of Lemma 5.3.7. Let $h_2 = \lceil \log(2/\nu) \rceil$, and let h_1 be the largest integer such that $(1 - 2^{-h_2})^{h_1} \geq 1 - \gamma$. Thus,

$$\begin{aligned}(1-\gamma) \leq (1-2^{-h_2})^{h_1} &\leq (1-\gamma)/(1-2^{-h_2}) \\ &< (1-\gamma)(1+2^{1-h_2}) \\ &\leq (1-\gamma)(1+\nu) \\ &< 1-\gamma+\nu\end{aligned}$$

and $h_1 = O(2^{h_2})$.

Define

$$\mathcal{C}(x) = \bigwedge_{g_1=1}^{h_1} \bigvee_{g_2=1}^{h_2} x_{g_1,g_2}.$$

and $h = h_1 h_2 = O(h_2 2^{h_2}) = O\left(\frac{1}{\nu} \log\left(\frac{1}{\nu}\right)\right)$.

Thus $\mathbf{Pr}_{x \sim \mathbf{U}_h}[\mathcal{C}(x) = 0] = 1 - (1 - 2^{-h_2})^{h_1}$, and hence

$$\gamma - \nu \leq \mathbf{Pr}_{x \sim \mathbf{U}_h}[\mathcal{C}(x) = 0] \leq \gamma.$$

□

We now proceed to prove Lemma 5.3.3 and Lemma 5.3.5.

For convenience, define

$$f^i_{\mathrm{Ext}}(y) = \bigvee_{1\le j\le M} \bigwedge_{\ell\in P^i_j} y_\ell$$

where $i \in \{0,1\}^r$. Further, let

$$p_2 = (1-p_1)^B, \quad p_3 = (1-p_2)^M.$$

We record two easy claims.

Claim 5.3.8. *For any $i \in \{0,1\}^r, j \in \{0,1\}^m$, $\mathbf{Pr}_{y\sim\mathbf{Ber}(s,1-p_1)}[\bigwedge_{\ell\in P^i_j} y_\ell = 1] = (1-p_1)^B = p_2$.*

Claim 5.3.9. *For any $i \in \{0,1\}^r$, $\mathbf{Pr}_{y\sim\mathbf{Ber}(s,1-p_1)}[f^i_{\mathrm{Ext}}(y) = 0] = (1-p_2)^M = p_3$.*

We frequently use the following inequality.

Claim 5.3.10. *For any $n > 1$ and $0 \le x \le n$, we have*

$$e^{-x}\left(1-\frac{x^2}{n}\right) \le \left(1-\frac{x}{n}\right)^n \le e^{-x}.$$

We also frequently use the following bounds.

Claim 5.3.11. *The following inequalities hold: Let $\epsilon_2 = \epsilon_1/2$. Then,*

1. $\frac{\ln R-\ln\ln 2}{M}\left(1-\frac{1}{B^{\epsilon_2}}\right) \le p_2 \le \frac{\ln R-\ln\ln 2}{M}\left(1+\frac{1}{B^{\epsilon_2}}\right) \le \frac{r}{M}$.
2. $\frac{1}{2R} \le \left(\frac{\ln 2}{R}\right)\left(1-\frac{2r}{B^{\epsilon_2}}\right) \le p_3 \le \left(\frac{\ln 2}{R}\right)\left(1+\frac{r}{B^{\epsilon_2}}\right) \le \frac{0.9}{R}$.

Proof. We have,

$$\begin{aligned} p_2 = (1-p_1)^B &\ge (1-\gamma)^B \ge e^{-\gamma B}(1-\gamma^2 B) && \text{(by Claim 5.3.10)} \\ &\ge \frac{\ln R-\ln\ln 2}{M}\left(1-\frac{r^2}{B}\right) && \text{(since } \gamma < (\ln M)/B < r/B\text{)} \end{aligned}$$

We now upper bound p_2. We have,

$$\begin{aligned}
p_2 &\le (1-\gamma(1-B^{-\epsilon_1}))^B \le e^{-\gamma B(1-B^{-\epsilon_1})} && \text{(by Claim 5.3.10)}\\
&< \left(\frac{\ln R - \ln\ln 2}{M}\right) M^{B^{-\epsilon_1}} < \left(\frac{\ln R - \ln\ln 2}{M}\right) e^{\delta r B^{-\epsilon_1}}\\
&\le \frac{\ln R - \ln\ln 2}{M}\left(1+\frac{r}{B^{\epsilon_1}}\right)
\end{aligned}$$

Thus,

$$\frac{\ln R - \ln\ln 2}{M}\left(1-\frac{1}{B^{\epsilon_2}}\right) \le p_2 \le \frac{\ln R - \ln\ln 2}{M}\left(1+\frac{1}{B^{\epsilon_2}}\right),$$

since $\epsilon_2 = \epsilon_1/2$.

Estimating similarly as above, we have

$$\begin{aligned}
p_3 &= (1-p_2)^M\\
&\ge \left(1-\left(\frac{\ln R - \ln\ln 2}{M}\right)\left(1+\frac{1}{B^{\epsilon_2}}\right)\right)^M\\
&\ge \left(1-\frac{(\ln R - \ln\ln 2)^2}{M}\left(1+\frac{1}{B^{\epsilon_2}}\right)^2\right)\left(\frac{\ln 2}{R}\right) e^{\frac{-(\ln R - \ln\ln 2)}{B^{\epsilon_2}}} && \text{(by Claim 5.3.10)}\\
&\ge \left(1-\frac{2r^2}{M}\right)\left(\frac{\ln 2}{R}\right) e^{-r/B^{\epsilon_2}}\\
&\ge \left(1-\frac{2r^2}{M}\right)\left(\frac{\ln 2}{R}\right)\left(1-\frac{r}{B^{\epsilon_2}}\right)\\
&\ge \left(1-\frac{2r}{B^{\epsilon_2}}\right)\left(\frac{\ln 2}{R}\right).
\end{aligned}$$

Finally, we have

$$\begin{aligned} p_3 &\leq \left(1 - \left(\frac{\ln R - \ln\ln 2}{M}\right)\left(1 - \frac{1}{B^{\epsilon_2}}\right)\right)^M \\ &\leq \left(\frac{\ln 2}{R}\right)^{1-B^{-\epsilon_2}} && \text{(by Claim 5.3.10)} \\ &\leq \left(\frac{\ln 2}{R}\right) 2^{r/B^{\epsilon_2}} \leq \left(\frac{\ln 2}{R}\right)\left(1 + \frac{r}{B^{\epsilon_2}}\right). \end{aligned}$$

Thus,

$$\left(\frac{\ln 2}{R}\right)\left(1 - \frac{2r}{B^{\epsilon_2}}\right) \leq p_3 \leq \left(\frac{\ln 2}{R}\right)^{1-\frac{r}{B}} \leq \left(\frac{\ln 2}{R}\right)\left(1 + \frac{r}{B^{\epsilon_2}}\right).$$

□

5.3.2 Proof of Lemma 5.3.3 : Bound on Influence of Coalitions on f_{Ext}

We now proceed to bound the influence of coalitions of variables on f_{Ext}.

Claim 5.3.12. *For any $i \in \{0,1\}^r$ and $q \leq s^{1-\delta}$, $\mathbf{I}_{q,\mathbf{Ber}(s,1-p_1)}(f^i_{\text{Ext}}) \leq \frac{1}{R}$.*

Proof. Let Q be any set of variables of size q, $q \leq s^{1-\delta}$. There are at most q blocks of P^i which contain a variable from Q. By Claim 5.3.8, it follows that the probability that for a y sampled from $\mathbf{Ber}(s, 1-p_1)$, there is no ANn gate at depth 1 in f^i_{Ext} which outputs 1 is at most

$$\begin{aligned} (1-p_2)^{M-q} &\leq p_3^{1-\frac{s^{1-\delta}}{M}} \\ &\leq p_3(2R)^{\frac{s^{1-\delta}}{M}} && \text{(since } p_3 > 1/(2R) \text{ by Claim 5.3.11)} \\ &\leq p_3 e^{r/M^{\delta/2}} && \text{(since } s = M^{1+\delta_1} < M^{1+\frac{\delta}{2}}/2) \\ &< \frac{1}{R} && \text{(since } p_3 < 0.9/R \text{ by Claim 5.3.11)} \end{aligned}$$

Thus the influence of Q is bounded by $\frac{1}{R}$. □

Definition 5.3.13. *For any $i \in \{0,1\}^r$ and $j \in \{0,1\}^m$, define a block P^i_j to be bad with respect*

to a subset of variables Q if $|P^i_j \cap Q| \geq 2\epsilon B$. Further call a partition P^i bad with respect to Q if it has a block which is bad. Otherwise, P^i is good.

Claim 5.3.14. *Consider any subset of variables Q of size q. If $q \leq s^{1-\delta}$, then there are less than KM bad partitions with respect to Q.*

Proof. Suppose to the contrary that there are at least KM bad partitions. It follows by an averaging argument that there exists $j \in \{0,1\}^m$ such that the number of bad blocks among the $\{P^i_j : i \in \{0,1\}^r\}$ is at least K. Define the function $\mathrm{Ext}_j(x,y) = \mathrm{Ext}(x,y) \oplus j$. Observe that Ext_j is a seeded extractor for min-entropy k with error ϵ.

Let $\mathcal{N}_j(x)$ denote the set of neighbours of x in the graph corresponding to Ext_j. It follows that $|\{|\mathcal{N}_j(x) \cap Q| \geq 2\epsilon B|\}| \geq K$. We note that $q/M = s^{1-\delta}/M = (MB)^{1-\delta}/M < 1/M^{\delta/19} < \epsilon$, since $\epsilon > 1/M^{\delta_1} = 1/M^{\delta/20} > 1/M^{\delta/19}$. Thus, we have

$$|\{|\mathcal{N}_j(x) \cap Q| \geq (\epsilon + \mu_Q)B\}| \geq K,$$

where $\mu_Q = q/M$. However this contradicts Theorem 2.4.1. Thus the number of bad blocks is bounded by KM. □

Claim 5.3.15. *Let P^i be a partition that is good with respect to a subset of variables Q, $|Q| = q$. If $q \leq s^{1-\delta}$, then $\mathbf{I}_{Q,\mathbf{Ber}(s,1-p_1)}(f^i_{\mathrm{Ext}}) \leq \frac{q}{2Rs^{1-\delta}}$.*

Proof. We note that there are at least $M - q$ blocks in P^i that do not have any variables from Q. Each of the remaining blocks have at most $2\epsilon B$ variables from Q. An assignment of x leaves f^i_{Ext} undetermined only if: (a) there is no ANn gate at depth 1 in f^i_{Ext} which outputs 1 and (b) There is at least one block with a variable from Q such that the non-Q variables are all set to 1. These two events are independent. Further, by Claim 5.3.12, the probability of (a) is bounded by $1/R$. We now bound the probability of (b). If there are h variables of Q in P^i_j, the probability that the

non-Q variables are all 1's is exactly $(1-p_1)^{B-h}$. Thus the probability of event (b) is bounded by

$$\begin{aligned}
q(1-p_1)^{B(1-2\epsilon)} &= qp_2^{1-2\epsilon} \\
&\leq \frac{qr}{M^{1-2\epsilon}} && \text{(since } p_2 < r/M \text{ by Claim 5.3.11)} \\
&= \frac{qr}{M^{1-\frac{\delta}{2}}} && \text{(since } \epsilon < \delta/4) \\
&< \frac{q}{M^{1-\frac{2\delta}{3}}} && \text{(using } r = M^{o(1)}) \\
&< \frac{q}{2s^{1-\delta}} && \text{(since } s = M^{1+\delta_1} < M^{1+\frac{\delta}{4}}).
\end{aligned}$$

□

Thus for any $q \leq s^{1-\delta}$,

$$\mathbf{I}_{q,\mathbf{Ber}(s,1-p_1)}(f_{\mathrm{Ext}}) \leq \frac{KM}{R} + \frac{q}{2s^{1-\delta}} = \frac{1}{R^{1-3\delta}} + \frac{q}{2s^{1-\delta}} < \frac{q}{s^{1-\delta}}.$$

□

5.3.3 Proof of Lemma 5.3.5: Bound on the Bias of f_{Ext}

We now proceed to show that f_{Ext} is almost balanced. For ease of presentation, we slightly abuse notation and relabel the partitions in Construction 1 as $P^1, \ldots, P^R$, where for any $i \in [R]$, P^i corresponds to the partition P^{v_i} with v_i being the r bit string for the integer $i-1$.

Claim 5.3.16. *There exists a small constant $\epsilon_3 > 0$ such that for any $i \in \{0,1\}^r$, $\mathbf{Pr}_{y\sim\mathbf{Ber}(s,1-p_1)}[f^i_{\mathrm{Ext}}(y) = 1] = 1 - \frac{\alpha}{R}$, where $1 - \frac{1}{B^{\epsilon_3}} \leq \frac{\alpha}{\ln 2} \leq 1 + \frac{1}{B^{\epsilon_3}}$.*

Proof. nirectly follows from Claim 5.3.11. □

We now estimate the probability $\mathbf{Pr}_{y\sim\mathbf{Ber}(s,1-p_1)}[f_{\mathrm{Ext}}(y) = 0]$. This is not direct since the f^i_{Ext}'s are on the same set of variables, and can be correlated in general. Towards estimating this, we introduce some definitions.

Definition 5.3.17. *Let P^i, P^j be two equi-partitions of $[s]$ with blocks of size B. Then (P^i, P^j) is said to be pairwise-good if the size of the intersection of any block of P^i and any block of P^j is at most $0.9B$.*

Definition 5.3.18. *Let $P^1, \ldots, P^R$ be equi-partitions of $[s]$ with blocks of size B. A collection of partitions $\mathcal{P} = \{P^1, \ldots, P^R\}$ is pairwise-good if for any distinct $i, j \in \{0,1\}^r$, (P^i, P^j) is pairwise-good.*

Lemma 5.3.19. *If $\mathcal{P}$ is pairwise-good, then $|\mathbf{E}_{y \sim \mathbf{Ber}(s,1-p_1)}[f_{\mathrm{Ext}}(y)] - \frac{1}{2}| \leq \frac{1}{B^{\Omega(1)}}$.*

Lemma 5.3.20. *The set of partitions $\mathcal{P} = \{P^1, \ldots, P^R\}$ in Construction 1 is pairwise-good.*

It is clear that the above two lemmas directly imply that $|\mathbf{E}_{y \sim \mathbf{Ber}(s,1-p_1)}[f_{\mathrm{Ext}}(y)] - \frac{1}{2}| \leq \frac{1}{B^{\Omega(1)}}$.

Proof of Lemma 5.3.20. Let $P^{i_1}_{j_1}$ and $P^{i_2}_{j_2}$ be any two blocks such that $i_1 \neq i_2$. We need to prove that $|P^{i_1}_{j_1} \cap P^{i_2}_{j_2}| \leq 0.9B$. Recall that $P^{i_1}_{j_1} = \{(z, \mathrm{Ext}(i_1, z) \oplus j_1) : z \in \{0,1\}^b\}$, and similarly $P^{i_2}_{j_2} = \{(z, \mathrm{Ext}(i_2, z) \oplus j_2) : z \in \{0,1\}^b\}$. The bound on $|P^{i_1}_{j_1} \cap P^{i_2}_{j_2}|$ now directly follows from the fact that Ext is a $\frac{1}{10}$-shift-design extractor. □

Proof of Lemma 5.3.19. Let $\mathcal{P} = \{P^1, \ldots, P^R\}$ be pairwise-good.

Recall that

$$p_3 = \mathbf{Pr}_{y \sim \mathbf{Ber}(s,1-p_1)}[f^i_{\mathrm{Ext}}(y) = 0] = \frac{\alpha}{R}.$$

Let y be sampled from $\mathbf{Ber}(s, 1-p_1)$. Let E_i be the event $f^i_{\mathrm{Ext}}(y) = 0$. We have,

$$p = \mathbf{Pr}_{y \sim \mathbf{Ber}(s,1-p_1)}[f_{\mathrm{Ext}}(y) = 0] = \mathbf{Pr}\left[\bigvee_{1 \leq i \leq R} E_i\right].$$

For $1 \leq c \leq R$, let

$$S_c = \sum_{1 \leq i_1 < \ldots < i_c \leq R} \mathbf{Pr}\left[\bigwedge_{1 \leq g \leq c} E_{i_g}\right].$$

Using the Bonferroni inequalities, it follows that for any even $a \in [R]$,

$$\sum_{c=1}^{a}(-1)^{(c-1)}S_c \leq p \leq \sum_{c=1}^{a+1}(-1)^{(c-1)}S_c. \tag{5.2}$$

Towards proving a tight bound on p using (5.2), we prove the following lemma.

Lemma 5.3.21. *There exist constants $\beta_1, \beta_2 > 0$ such that for any $c \leq s^{\beta_1}$, and arbitrary $1 \leq i_1 < \ldots < i_c \leq R$, the following holds:*

$$\left(\frac{\alpha}{R}\right)^c \leq \mathbf{Pr}\left[\bigwedge_{1\leq g\leq c} E_{i_g}\right] \leq \left(\frac{\alpha}{R}\right)^c \left(1 + \frac{1}{M^{\beta_2}}\right).$$

To prove the above lemma, we recall Janson's inequality [Jan90, BS89]. We follow the presentation in [AS92].

Theorem 5.3.22 (Janson's Inequality [Jan90, BS89, AS92])**.** *Let Ω be a finite universal set and let $\mathcal{O}$ be a random subset of Ω constructed by picking each $h \in \Omega$ independently with probability p_h. Let $Q_1, \ldots, Q_\ell$ be arbitrary subsets of Ω, and let $\mathcal{E}_i$ be the event $Q_i \subseteq \mathcal{O}$. Define*

$$\Delta = \sum_{i<j : Q_i \cap Q_j \neq \emptyset} \mathbf{Pr}\left[\mathcal{E}_i \wedge \mathcal{E}_j\right], \qquad n = \prod_{i=1}^{\ell} \mathbf{Pr}\left[\overline{\mathcal{E}_i}\right].$$

Assume that $\mathbf{Pr}[\mathcal{E}_i] \leq \tau$ for all $i \in [\ell]$. Then

$$n \leq \mathbf{Pr}\left[\bigwedge \overline{\mathcal{E}_i}\right] \leq n e^{\frac{\Delta}{1-\tau}}.$$

□

Proof of Lemma 5.3.21. We set $\beta_1 = 1/90$ with foresight. Without loss of generality suppose $i_g = g$ for $g \in [c]$. We use Janson's inequality with $\Omega = [s]$, and $\mathcal{O}$ constructed by picking each $h \in [s]$ with probability $1 - p_1$. Further let $\mathcal{E}_{i,j}$ be the event that $P_j^i \subseteq \mathcal{O}$. Intuitively, $\mathcal{O}$ denotes the set

of coordinates in y that are set to 1 for a sample y from $\mathbf{Ber}(s, 1-p_1)$. With this interpretation, the event $f^i_{\mathrm{Ext}}(y) = 0$ exactly corresponds to the event $\bigwedge_{1\le j\le M} \overline{\mathcal{E}_{i,j}}$. Thus, we have

$$\mathbf{Pr}\left[\bigwedge_{1\le g\le c} E_g\right] = \mathbf{Pr}\left[\bigwedge_{i\in[c], j\in\{0,1\}^m} \overline{\mathcal{E}_{i,j}}\right].$$

We now estimate n, Δ, γ to apply Janson's inequality. For any $i \in [c], j \in \{0,1\}^m$, we have $\mathbf{Pr}[\mathcal{E}_{i,j}] = \mathbf{Pr}[P^i_j \subseteq \mathcal{O}] = (1-p_1)^B = p_2$. Note that $\tau = p_2 < \frac{1}{2}$. Further

$$n = \prod_{i\in[c], j\in\{0,1\}^m} \mathbf{Pr}\left[\overline{\mathcal{E}_{i,j}}\right] = (1-p_2)^{Mc} = p_3^c = \left(\frac{\alpha}{R}\right)^c.$$

Finally, we have

$$\Delta = \sum_{i_1<i_2\in[c], j_1, j_2\in\{0,1\}^m : P^{i_1}_{j_1}\cap P^{i_1}_{j_1}\neq\emptyset} \mathbf{Pr}[\mathcal{E}_{i_1,j_1} \wedge \mathcal{E}_{i_2,j_2}]$$

We observe that any P^i_j can intersect at most B blocks of another partition $P^{i'}$. Thus, the total number of blocks that intersect between two partitions P^i and P^j is bounded by $MB = s$. Further, recall that $\mathcal{P}$ is pairwise-good. Thus it follows that for any distinct $i_1, i_2 \in [c]$, and $j_1, j_2 \in \{0,1\}^m$, $|P^{i_1}_{j_1} \cap P^{i_2}_{j_2}| \le 0.9B$. Thus, $|P^{i_1}_{j_1} \cup P^{i_2}_{j_2}| \ge 1.1B$ and hence for any $i_1 < i_2 \in [c], j_1, j_2 \in \{0,1\}^m$,

$$\mathbf{Pr}[\mathcal{E}_{i_1,j_1} \wedge \mathcal{E}_{i_2,j_2}] \le (1-p_1)^{\frac{11B}{10}} = p_2^{\frac{11}{10}}.$$

By Claim 5.3.11, $p_2 \le \frac{r}{M}$. Thus,

$$\Delta \le \binom{c}{2} s p_2^{\frac{11}{10}} < \frac{s^{1+2\beta_1} r^2}{M^{\frac{11}{10}}} = \frac{(MB)^{1+2\beta_1} r^2}{M^{\frac{11}{10}}} = \frac{B^{1+2\beta_1} r^2}{M^{\frac{1}{10}-2\beta_1}} = \frac{M^{\delta_1(1+2\beta_1)} r^2}{M^{\frac{1}{10}-2\beta_1}} < \frac{r^2}{M^{\frac{1}{20}-3\beta_1}}.$$

Recall $\beta_1 = 1/90$. It follows that

$$\Delta < M^{-\beta'},$$

where $\beta' = 1/70$.

Invoking Janson's inequality, we have

$$\left(\frac{\alpha}{R}\right)^c \leq \mathbf{Pr}\left[\bigwedge_{1\leq g\leq c} E_g\right] \leq \left(\frac{\alpha}{R}\right)^c e^{2M^{-\beta'}} \leq \left(1+\frac{3}{M^{\beta'}}\right)\left(\frac{\alpha}{R}\right)^c.$$

This concludes the proof. □

Fix $a = s^{\beta_3}$ (assume that a is even), $\beta_3 = \min\{\beta_1/2, \beta_2/1000\}$, where β_1, β_2 are the constants in Lemma 5.3.21.

The following lemma combined with (5.2) proves a tight bound on p (recall that $p = \mathbf{Pr}_{y\sim\mathbf{Ber}(s,1-p_1)}[f_{\mathrm{Ext}}(y) = 0]$).

Claim 5.3.23. $e^{-\alpha} - \frac{1}{M^{\beta_2/2}} \leq \sum_{c=1}^{a}(-1)^{c-1}S_c < \sum_{c=1}^{a+1}(-1)^{c-1}S_c \leq e^{-\alpha} + \frac{1}{M^{\beta_2/2}}$.

Proof. For any $c \leq a+1$, using Lemma 5.3.21, we have

$$\binom{R}{c}\left(\frac{\alpha}{R}\right)^c \leq S_c \leq \binom{R}{c}\left(\frac{\alpha}{R}\right)^c\left(1+\frac{1}{M^{\beta_2}}\right).$$

We have,

$$\begin{aligned}\binom{R}{c}\left(\frac{\alpha}{R}\right)^c &\leq \frac{R^c}{c!}\frac{\alpha^c}{R^c}\\ &= \frac{\alpha^c}{c!}\end{aligned}$$

and

$$\begin{aligned}\binom{R}{c}\left(\frac{\alpha}{R}\right)^c &= \frac{R(R-1)\dots(R-c+1)}{R^c}\frac{\alpha^c}{c!}\\ &\geq \left(1-\frac{a^2}{R}\right)\frac{\alpha^c}{c!} && \text{(by Weierstrass product inequality)}\\ &\geq \left(1-\frac{1}{R^{1-\beta_2}}\right)\frac{\alpha^c}{c!}\end{aligned}$$

by our choice of a.

Thus, for any $c \leq a$, we have

$$\left|S_c - \frac{\alpha^c}{c!}\right| \leq \frac{1}{M^{\beta_2}} \tag{5.3}$$

It also follows that

$$S_{a+1} \leq \frac{1}{a!} + \frac{1}{M^{\beta_2}} < \frac{2}{M^{\beta_2}}, \tag{5.4}$$

using $a = s^{\beta_3}$.

Finally, by the classical Taylor's theorem, we have

$$\left|e^{-\alpha} - \sum_{c=1}^{a}(-1)^{c-1}\frac{\alpha^c}{c!}\right| < \frac{1}{a!} < \frac{1}{M^{\beta_2}}. \tag{5.5}$$

Claim 5.3.23 is now direct from the inequalities (5.3), (5.4), (5.5) and the fact that $aM^{-\beta_2} \leq M^{-\beta_2/2}$. □

The next claim is a restatement of Lemma 5.3.19.

Claim 5.3.24. *$|p - \frac{1}{2}| \leq B^{-\Omega(1)}$, where $p = \mathbf{Pr}_{y\sim\mathbf{Ber}(s,1-p_1)}[f_{\mathrm{Ext}}(y) = 0]$.*

Proof. Using (5.2) and Claim 5.3.23, we have

$$|p - e^{-\alpha}| \leq \frac{1}{M^{\beta_2/2}}.$$

Recall that from Claim 5.3.16, we have

$$\ln 2 \left(1 - \frac{1}{B^{\epsilon_3}}\right) \leq \alpha \leq \ln 2 \left(1 + \frac{1}{B^{\epsilon_3}}\right).$$

Thus,

$$\left|e^{-\alpha} - \frac{1}{2}\right| \leq \frac{2}{B^{\epsilon_3}}$$

and hence, we have

$$\left|p - \frac{1}{2}\right| \leq \frac{3}{B^{\epsilon_3}}.$$

□

□

Chapter 6

Two-Source Extractors and Ramsey Graphs

[1] An extractor Ext : $\{0,1\}^n \to \{0,1\}^m$ is a deterministic function that takes input from a weak source with sufficient min-entropy and produces nearly uniform bits. Unfortunately, it is impossible to extract even 1 bit for sources with min-entropy $n-1$. To circumvent this difficulty, Santha and Vazirani [SV86], and Chor and Goldreich [CG88] suggested the problem of designing extractors for two or more independent sources, each with sufficient min-entropy. When the extractor has access to just two sources, it is called a two-source extractor. An efficient two-source extractor could be quite useful in practice, if just two independent sources of entropy can be found.

Definition 6.0.1 (Two-source extractor)**.** *A function* Ext : $\{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ *is called a* (k,ϵ)*-two-source extractor if for any independent* (n,k)*-sources* $\mathbf{X}$ *and* $\mathbf{Y}$*, we have*

$$|\text{Ext}(\mathbf{X},\mathbf{Y}) - \mathbf{U}_m| \leq \epsilon.$$

Note that for $m = 1$, this corresponds to an $N \times N$ matrix with entries in $\{0,1\}$ such that

[1] parts of this chapter have been previously published [CZ16a, CL16a]

every $K \times K$ submatrix has $1/2 \pm \varepsilon$ fraction of 1's, where $N = 2^n$ and $K = 2^k$.

6.1 Prior Work and Our results

Chor and Goldreich [CG88] used Lindsey's Lemma to show that the inner-product function is a 2-source extractor for min-entropy more than $n/2$. However, no further progress was made for around 20 years, when Bourgain [Bou05b] broke the "half-barrier" for min-entropy, and constructed a 2-source extractor for min-entropy $0.499n$. This remains the best known result prior to this work. Bourgain's extractor was based on breakthroughs made in the area of additive combinatorics.

Raz [Raz05] obtained an improvement in terms of total min-entropy, and constructed 2-source extractors requiring one source with min-entropy more than $n/2$ and the other source with min-entropy $O(\log n)$. A different line of work investigated a weaker problem of designing dispersers for two independent sources due to its connection with Ramsey graphs. We discuss this in Section 6.2.

The lack of progress on constructing two-source extractors motivated researchers to use more than two sources with the best known result due to Li [Li13a], where he showed how to extract from 3 sources, each with polylogarithmic min-entropy. We discuss this line of work in more detail in Chapter 7. Thus, in summary, despite much attention and progress over the last 30 years, it remained open to explicitly construct two-source extractors for min-entropy rate significantly smaller than $1/2$.

In joint work with Zuckerman [CZ16a], we construct an explicit two-source extractor for polylogarithmic min-entropy.

Theorem 14. *There exists a constant $C > 0$ such that for all $n \in \mathbb{N}$, there exists a polynomial time computable construction of a 2-source extractor $\mathrm{2Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ for min-entropy at least $\log^C(n)$ and error $n^{-\Omega(1)}$.*

The constant C in the above theorem can be taken to be 75. This was improved to $C = 19$ by Meka [Mek15]. Subsequently in joint work with Li [CL16a], we improve this to 14 (see Section

6.5 for more details). By an argument of Barak [Rao09b], every 2-source extractor is also a strong 2-source extractor with similar parameters. Thus the extractor 2Ext in Theorem 14 is also a strong 2-source extractor.

An open problem here is to improve the error to negligible since this is important useful for applications in cryptography and distributed computing. For example, several researchers have studied whether cryptographic or distributed computing protocols can be implemented if the players' randomness is defective [DO03, GSV05, KLRZ08, KLR09]. Kalai et al. [KLRZ08] used C-source extractors to build network extractor protocols, which allow players to extract private randomness in a network with Byzantine faults. A better 2-source extractor with negligible error would improve some of those constructions. Kalai, Li, and Rao [KLR09] showed how to construct a 2-source extractor under computational assumptions, and used it to improve earlier network extractors in the computational setting; however, their protocols rely on computational assumptions beyond the 2-source extractor, so it would not be clear how to match their results without assumptions.

If we allow the 2-source extractor to run in time $\text{poly}(n, 1/\epsilon)$, then our technique in fact generalizes to obtain arbitrary error ϵ. In particular, we have the following theorem.

Theorem 15. *There exists a constant $C > 0$ such that for all $n \in \mathbb{N}$ and any $\epsilon > 0$, there exists a 2-source extractor $\text{2Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ computable in time $\text{poly}(n, 1/\epsilon)$ for min-entropy at least $\log^C(n/\epsilon)$ and error ϵ.*

Recently, Li [Li15a] extended the construction in [CZ16a] to achieve an explicit strong 2-extractor with output length k^α bits, for some small constant α. By our observation above, this immediately implies a 2-source extractor for min-entropy $k \geq \log^{C'} n$, for some large enough constant C', with output length $\Omega(k)$; in fact, the output can be k bits.

6.2 Ramsey Graphs

Definition 6.2.1 (Ramsey graphs)**.** *A graph on N vertices is called a K-Ramsey graph if does not contain any independent set or clique of size K.*

It was shown by Erdös in one of the first applications of the probabilistic method that there exists K-Ramsey graphs for $K = 2\log N$. By explicit, we mean a polynomial-time algorithm that determines whether there is an edge between two nodes, i.e., the running time should be polylogarithmic in the number of nodes.

Frankl and Wilson [FW81] used intersection theorems to construct K-Ramsey graphs on N vertices, with $K = 2^{O(\sqrt{\log N \log\log N})}$. This remained the best known construction for a long time, with many other constructions [Alo98, Gro00, Bar06] achieving the same bound. Gopalan [Gop14] explained why approaches were stuck at this bound, showing that apart from [Bar06], all other constructions can be seen as derived from low-degree symmetric representations of the OR function. Finally, subsequent works by Barak et al. [BKS$^+$10, BRSW12] obtained a significant improvement and gave explicit constructions of K-Ramsey graphs, with $K = 2^{2^{\log^{1-\alpha}(\log N)}}$, for some absolute constant α.

We also define a harder variant of Ramsey graphs.

Definition 6.2.2 (Bipartite Ramsey graph)**.** *A bipartite graph with N left vertices and N right vertices is called a bipartite K-Ramsey graph if it does not contain any complete $K \times K$-bipartite sub-graph or empty $K \times K$ sub-graph.*

Explicit bipartite K-Ramsey graphs were known for $K = \sqrt{N}$ based on the Hadamard matrix. This was slightly improved to $o(\sqrt{N})$ by Pudlak and Rődl [PR04], and the results of [BKS$^+$10,BRSW12] in fact constructed bipartite K-Ramsey graphs, and hence achieved the bounds as mentioned above.

The following lemma is easy to obtain (see e.g.,[BRSW12]).

Lemma 6.2.3. *Suppose that for all $n \in \mathbb{N}$ there exists a polynomial time computable 2-source extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ *for min-entropy k and error $\epsilon < 1/2$. Let $N = 2^n$ and $K = 2^k$. Then there exists an explicit construction of a bipartite K-Ramsey on N vertices.*

Thus, Theorem 14 implies the following.

Theorem 16. *There exists a constant $C > 0$ such that for all large enough $n \in \mathbb{N}$, there exists an explicit construction of a bipartite K-Ramsey graph on $2N$ vertices, where $N = 2^n$ and $K = 2^{(\log\log N)^C}$.*

The constant C in [CZ16a] can be taken to be 75. This was improved to $C = 11$ by Meka [Mek15], and subsequently improved in [CL16a] to 8 (see Section 6.5).

Given any bipartite K-Ramsey graph, a simple reduction gives a $K/2$-Ramsey graph on N vertices [BKS+10]. As an immediate corollary, we have explicit constructions of Ramsey graphs with the same bound.

Corollary 6.2.4. *There exists a constant $C > 0$ such that for all large enough $n \in \mathbb{N}$, there exists an explicit construction of a K-Ramsey graph on N vertices, where $N = 2^n$ and $K = 2^{(\log\log N)^C}$.*

Independent work: In independent work[2], Cohen [Coh16c] used the challenge-response mechanism introduced in [BKS+10] with new advances in constructions of extractors to obtain a two-source disperser for polylogarithmic min-entropy. Using this, he obtained explicit constructions of bipartite-Ramsey graphs with $K = 2^{(\log\log N)^{O(1)}}$, which matches our result and thus provides an alternate construction.

6.3 An Outline of Our 2-Source Extractor Construction

To motivate our construction, first, let's try to build a 1-source extractor (even though we know it is impossible). Let $\mathbf{X}$ be an (n, k)-source, where $k = \text{polylog}(n)$. Let Ext be a strong seeded extractor designed to extract 1 bit from min-entropy k with error ϵ. Since, for $(1-\epsilon)$-fraction of the seeds, the extractor output is close to uniform, a natural idea is to do the following: cycle over all the seeds of Ext and concatenate the outputs to obtain a D-bit string $\mathbf{Z}$ where most individual bits are close to uniform. Note that since the seed length of Ext is $O(\log n)$, $D = \text{poly}(n)$. At this point, we might hope to take majority of these D bits of $\mathbf{Z}$ to obtain a bit is close to uniform. However,

[2]Cohen's work appeared before our work on 2-source exttractors [CZ16a]. When his paper appeared, we had an outline of the proof but had not filled in the details.

the output of Ext with different seeds may be correlated in arbitrary ways (even if individually the bits are close to uniform), so this approach doesn't work.

We try to fix this approach by introducing some independence among the uniform bits. For example, if we obtain a source $\mathbf{Z}$ such that $D - D^{0.49}$ bits are uniform, and further these bits are (almost) constant-wise independent, then it is known that the majority function can extract an almost-uniform bit (see Lemma 7.3.3). In an attempt to obtain such a source, we use explicit t-non-malleable extractors from Chapter 4. Let nmExt be a (t, k, ϵ)-non-malleable extractor that outputs 1 bit with seed-length d, and let $D = 2^d$. We show in Lemma 6.4.3, that there exists a large subset of seeds $S \subset \{0,1\}^d$, $|S| \geq (1 - O(\sqrt{\epsilon}))D$, such that for any t distinct seeds $s_1, \ldots, s_t$ in S, $|\text{nmExt}(\mathbf{X}, s_1), \ldots, \text{nmExt}(\mathbf{X}, s_t) - U_t| \leq O(t\sqrt{\epsilon})$. Thus, we could use our earlier idea of cycling through all seeds, but now using an explicit non-malleable extractor instead of a strong-seeded extractor. We use the explicit t-non-malleable extractor constructed in Chapter 4 (see Theorem 1)). This construction requires min-entropy $k = \Omega(t \log^2(n/\epsilon))$ and seed-length $d = O(t^2 \log^2(n/\epsilon))$. Thus, we could cycle over all the seeds of nmExt, and produce a string $\mathbf{Z}$ of length $D = 2^{O(t^2 \log^2(n/\epsilon))}$, such that the i'th bit of $\mathbf{Z}$, $\mathbf{Z}_i = \text{nmExt}(\mathbf{X}, i)$. Further, except for at most $O(\sqrt{\epsilon}D)$ bits in $\mathbf{Z}$, the remaining bits in $\mathbf{Z}$ follow a $(t, O(t\sqrt{\epsilon}))$-wise independent distribution. We could now try to set parameters such that the majority function extracts a bit from $\mathbf{Z}$. However, it is easy to check that $\sqrt{\epsilon}D > D^{1-\delta}$, for any constant $\delta > 0$. Since the majority function can handle at most $\sqrt{D}$ bad bits, this idea fails.

Our next idea is to look for functions that can handle larger number of "bad bits" to extract from $\mathbf{Z}$. This exactly corresponds to the notion of resilient functions studied in Chapter 5 and we note that $\mathbf{Z}$ is non-oblivious bit-fixing sources. Thus, our idea is to use the explicit $(\log(D))^{O(1)}$-independent $(D^{1-\delta}, D^{-\Omega(1)})$-resilient functions from Theorem 12 in Chapter 5.

Recall that $\mathbf{Z} = \text{nmExt}(\mathbf{X}, 1) \circ \ldots \circ \text{nmExt}(\mathbf{X}, D)$ is a (q, t, γ)-NOBF source on D bits, where $q = \sqrt{\epsilon}D, \gamma = O(\sqrt{\epsilon}t)$ and $D = 2^{O(t^2 \log^2(n/\epsilon))}$. We set $t = \log^{O(1)}(D)$, and thus we require $H_\infty(\mathbf{X}) = \log^{(O(1))}(n/\epsilon)$. As we observed before, $q > D^{1-\delta}$ for any $\delta > 0$. Thus, we cannot directly apply the resilient function f from Theorem 12 on $\mathbf{Z}$ to extract an almost bit. (A more important

issue in directly applying f to $\mathbf{Z}$ is that while using Lemma 5.2.1, we have to bound the term γD^t in the error, which is clearly greater than 1 for the current parameters.) We note that it is not surprising that f cannot extract from $\mathbf{Z}$ since we just used 1 source up to this point.

We now use the second independent source $\mathbf{Y}$ to sample a pseudorandom subset T of coordinates from $[D]$, $|T| = D' = n^{O(1)}$, such that the fraction of bad bits $\mathbf{Z}_T$ (the projection of $\mathbf{Z}$ to the coordinates in T) remains almost the same as that of $\mathbf{Z}$ (with high probability). A well known way of using a weak source to sample a pseudorandom subset was discovered by Zuckerman [Zuc97], and uses a seeded extractor, with the size of the sample being the total number of seeds and fraction of bad bits increases at most by the error of the extractor (with high probability). Thus using known optimal constructions of seeded extractors with seed-length $d' = O(\log(n/\epsilon'))$, we have $D' = (n/\epsilon')^{O(1)}$. Thus $\mathbf{Z}_t$ is (q, t, γ)-NOBF source on D' bits, where $q = (\sqrt{\epsilon} + \epsilon')D', \gamma = O(\sqrt{\epsilon}t)$. Further, the incurred error on applying f (from Theorem 12) on $\mathbf{Z}_t$ is $(D')^{-\Omega(1)} + \gamma(D')^t$ (using Lemma 5.2.1). By choosing δ to be a small enough constant, the term $\epsilon' D'$ can be made smaller than $(D')^{1-\delta}/2$. Further, by choosing ϵ small enough ($n^{-(\log n)^{O(1)}}$), we can ensure that $\sqrt{\epsilon}D' < (D')^{1-\delta}/2$ and $\gamma D' = (D')^{-\Omega(1)}$. This completes the description of our 2-source extractor.

6.4 Reduction to an NOBF Source

The main result in this section is a reduction from the problem of extracting from two independent (n, k)-sources to the task of extracting from a single (q, t, γ)-NOBF source on $n^{O(1)}$ bits. We formally state the reduction in the following theorem.

Theorem 6.4.1. *There exist constants $\delta, c' > 0$ such that for every $n, t > 0$ there exists a polynomial time computable function* reduce $: \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^D$, $D = n^{O(1)}$, *satisfying the following property: if $\mathbf{X}, \mathbf{Y}$ are independent (n, k)-sources with $k \geq c' t^4 \log^2 n$, then*

$$\mathbf{Pr}_{y \sim \mathbf{Y}}[\text{reduce}(\mathbf{X}, y) \text{ is a } (q, t, \gamma)\text{-NOBF source}] \geq 1 - n^{-\omega(1)}$$

where $q = D^{1-\delta}$ and $\gamma = 1/D^{t+1}$.

Li had earlier proved a similar theorem with $q = D/3$, and his methods would extend to achieve a similar bound as we achieve.

The δ we obtain in Theorem 6.4.1 is a small constant. Further, it can be shown that for our reduction method, it is not possible to achieve $\delta > 1/2$. Thus, we cannot use the majority function as the extractor for the resulting (q, t, γ)-NOBF source.

The reduction in Theorem 6.4.1 is based on explicit constructions of non-malleable extractors from Chapter 4.

In the following lemma, we reduce extracting from two independent sources to extracting from a (q, t, γ)-NOBF source using non-malleable extractors and seeded extractors in a black-box way. Theorem 6.4.1 then follows by plugging in explicit constructions of these components.

Lemma 6.4.2. *Let* $\mathrm{nmExt} : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}$ *be a* (t, k, ϵ_1)*-non-malleable extractor and let* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^{d_2} \to \{0,1\}^{d_1}$ *be a seeded extractor for min-entropy* $k/2$ *with error* ϵ_2*. Let* $\{0,1\}^{d_2} = \{s_1, \ldots, s_{D_2}\}$*,* $D_2 = 2^{d_2}$*. Suppose that* Ext *satisfies the property that for all* $y \in \{0,1\}^n$*,* $\mathrm{Ext}(y, s) \neq \mathrm{Ext}(y, s')$ *whenever* $s \neq s'$*. Define the function:*

$$\mathrm{reduce}(x, y) = \mathrm{nmExt}(x, \mathrm{Ext}(y, s_1)) \circ \ldots \circ \mathrm{nmExt}(x, \mathrm{Ext}(y, s_{D_2})).$$

If $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent* (n, k)*-sources, then*

$$\mathbf{Pr}_{y \sim \mathbf{Y}}[\mathrm{reduce}(\mathbf{X}, y) \text{ is a } (q, t, \gamma)\text{-NOBF source}] \geq 1 - n^{-\omega(1)},$$

where $q = (\sqrt{\epsilon_1} + \epsilon_2)D_2$ *and* $\gamma = 5t\sqrt{\epsilon_1}$*.*

We prove a lemma about t-non-malleable extractors from which Lemma 6.4.2 is easy to obtain.

Lemma 6.4.3. *Let* $\mathrm{nmExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}$ *be a* (t, k, ϵ)*-non-malleable extractor. Let* $\{0,1\}^d = \{s_1, \ldots, s_D\}$*,* $D = 2^d$*. Let* $\mathbf{X}$ *be any* (n, k)*-source. There exists a subset* $R \subseteq \{0,1\}^d$*,*

$|R| \geq (1-\sqrt{\epsilon})D$ *such that for any distinct* $r_1, \ldots, r_t \in R$,

$$(\mathrm{nmExt}(\mathbf{X}, r_1), \ldots, \mathrm{nmExt}(\mathbf{X}, r_t)) \approx_{5t\sqrt{\epsilon}} \mathbf{U}_t.$$

Proof. Let

$$\begin{aligned} BAD = \{r \in \{0,1\}^d : \exists \text{ distinct } r_1, \ldots, r_t \in \{0,1\}^d, \\ \forall i \in [t]\ r_i \neq r,\ \text{s.t } |(\mathrm{nmExt}(\mathbf{X}, r), \mathrm{nmExt}(\mathbf{X}, r_1), \ldots, \mathrm{nmExt}(\mathbf{X}, r_t)) - \\ (\mathbf{U}_1, \mathrm{nmExt}(\mathbf{X}, r_1), \ldots, \mathrm{nmExt}(\mathbf{X}, r_t))| > \sqrt{\epsilon}\} \end{aligned}$$

We define adversarial functions $f_1, \ldots, f_t$ as follows. For each $r \in BAD$, set $f_i(r) = r_i$, $i = 1, \ldots, t$ (the f_i's are defined arbitrarily for $r \notin BAD$, only ensuring that there are no fixed points). Let $\mathbf{Y}$ be uniform on $\{0,1\}^d$. It follows that

$$\begin{aligned} |(\mathrm{nmExt}(\mathbf{X}, \mathbf{Y}), \mathrm{nmExt}(\mathbf{X}, f_1(\mathbf{Y})), \ldots, \mathrm{nmExt}(\mathbf{X}, f_t(\mathbf{Y}))) - \\ (\mathbf{U}_1, \mathrm{nmExt}(\mathbf{X}, f_1(\mathbf{Y})), \ldots, \mathrm{nmExt}(\mathbf{X}, f_t(\mathbf{Y})))| \geq \frac{\sqrt{\epsilon}}{2^d}|BAD| \end{aligned}$$

Thus $|BAD| \leq \sqrt{\epsilon}2^d$ using the property that nmExt is a (k, t, ϵ)-non-malleable extractor. Define $R = \{0,1\}^d \setminus BAD$. Using Lemma 2.3.11, it follows that R satisfies the required property. □

Proof of Lemma 6.4.2. Let $R \subseteq \{0,1\}^{d_1}$ be such that for any distinct $r_1, \ldots, r_t \in R$,

$$(\mathrm{nmExt}(\mathbf{X}, r_1), \ldots, \mathrm{nmExt}(\mathbf{X}, r_t)) \approx_{5t\sqrt{\epsilon_1}} \mathbf{U}_t.$$

It follows by Lemma 6.4.3 that $|R| \geq (1-\sqrt{\epsilon_1})D_1$.

Define $\mathrm{Samp}(y) = \{\mathrm{Ext}(y, s_1), \ldots, \mathrm{Ext}(y, s_{D_2})\} \subset \{0,1\}^{d_1}$. Using Theorem 2.4.2, we have

$$\Pr_{y \sim \mathbf{Y}}[|\mathrm{Samp}(y) \cap R| \leq (1 - \sqrt{\epsilon_1} - \epsilon_2)D_2] \leq 2^{-k/2}. \tag{6.1}$$

Consider any y such that $|\text{Samp}(y) \cap R| \geq (1 - \sqrt{\epsilon_1} - \epsilon_2)D_2$, and let $\mathbf{Z}_y = \text{reduce}(\mathbf{X}, y)$. Since the output bits of nmExt corresponding to seeds in $\text{Samp}(y) \cap R$ are $(t, 5t\sqrt{\varepsilon_1})$-wise independent, we have that $\mathbf{Z}_y$ is a $((\sqrt{\epsilon_1} + \epsilon_2)D_2, t, 5t\sqrt{\varepsilon_1})$-NOBF source on D_2 bits.

Thus using (6.1), it follows that with probability at least $1 - 2^{-k/2}$ over $y \sim \mathbf{Y}$, $\text{reduce}(\mathbf{X}, y)$ is a $((\sqrt{\epsilon_1} + \epsilon_2)D_2, t, 5t\sqrt{\epsilon_1})$-NOBF source on D_2 bits. □

Proof of Theorem 6.4.1. We derive Theorem 6.4.1 from Lemma 6.4.2 by plugging in explicit non-malleable extractors and seeded extractors as follows:

1. Let $\text{nmExt} : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}$ be an explicit (t, k, ϵ_1)-non-malleable extractor from Theorem 1. Thus $d_1 = c_1 t^2 \log^2(n/\epsilon_1)$, for some constant c_1. Such an extractor exists as long as $k \geq \lambda_1 t \log^2(n/\epsilon_1)$ for some constant λ_1.

2. Let $\text{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^{d_1}$ be the extractor from Corollary 2.1.3 set to extract from min-entropy $k/2$ with error ϵ_2. Thus $d = c_2 \log(n/\epsilon_2)$ for some constant c_2. Let $D = 2^d = (n/\epsilon_2)^{c_2}$. Such an extractor exists as long as $k \geq 3d_1$.

3. We choose $\epsilon_1, \epsilon_2, \delta$ such that the following hold:

   - $(\sqrt{\epsilon_1} + \epsilon_2)D \leq D^{1-\delta}$.
   - $\sqrt{\epsilon_1} \leq 1/(5tD^{t+1})$.
   - $\delta' = \delta c_2 < 9/10$.

   To satisfy the above requirements, we pick ϵ_1, ϵ_2 as follows: Let $\epsilon_2 = 1/n^{C_2}$ where C_2 is fixed such that $\epsilon_2 D \leq D^{1-\delta}/2$. Thus, we need to ensure that
   $\epsilon_2 \leq 1/(2D^\delta)$. Substituting $D = (n/\epsilon_2)^{c_2}$ and simplifying, we have

$$\begin{aligned} \epsilon_2 &\leq \epsilon_2^{c_2\delta}/2n^{c_2\delta} \\ \text{i.e.,} \quad \epsilon_2^{1-c_2\delta} &\leq 1/2n^{c_2\delta} \\ \text{i.e.,} \quad \epsilon_2 &\leq 1/(2n)^{\delta'/(1-\delta')}. \end{aligned}$$

We note that $1 - \delta' > 1/10$. Thus, we can choose $C_2 = 10$.

We now set $\epsilon_1 = 1/n^{C_1 t}$, where we choose the constant C_1 such that $\sqrt{\epsilon_1} \leq 1/(5tD^{t+1})$. Simplifying, we have

$$\epsilon_1 \leq \frac{\epsilon_2^{2c_2(t+1)}}{25t^2 n^{2c_2(t+1)}} \leq \frac{1}{25t^2 n^{2c_2(C_2+1)(t+1)}} \leq \frac{1}{n^{23c_2(t+1)}}.$$

Thus, we can choose $C_1 = 24c_2$.

4. We note that for the above choice of parameters, nmExt and Ext indeed work for min-entropy $k \geq c' t^4 \log^2 n$, for some large constant c'.

5. Let $\{0,1\}^d = \{s_1, \ldots, s_D\}$.

Define the function:

$$\text{reduce}(x, y) = \text{nmExt}(x, \text{Ext}(y, s_1)) \circ \ldots \circ \text{nmExt}(x, \text{Ext}(y, s_D)).$$

Let $\mathbf{X}$ and $\mathbf{Y}$ be independent (n, k)-sources. By Lemma 6.4.2, it follows that

$$\mathbf{Pr}_{y \sim \mathbf{Y}}[\text{reduce}(\mathbf{X}, y) \text{ is a } (q, t, \gamma)\text{-NOBF source}] \geq 1 - n^{-\omega(1)},$$

where $q = (\sqrt{\epsilon_1} + \epsilon_2)D$ and $\gamma = 5t\sqrt{\epsilon_1}$. Theorem 6.4.1 now follows by our choice of parameters. □

6.5 Wrapping Up the Proofs of Theorem 13 and Theorem 14

Proof of Theorem 13. Let $f : \{0,1\}^n \to \{0,1\}$ be the explicit function constructed in Theorem 12 satisfying: For any $q > 0$, $t \geq c(\log n)^{18}$ (c is the constant from Theorem 12) and $\gamma \leq 1/n^{t+1}$,

- $\mathbf{I}_q(f) \leq q/n^{1-\frac{\delta}{2}}$
- For any (t, γ)-wise independent distribution $\mathcal{D}$, $\left|\mathbf{E}_{x \sim \mathcal{D}}[f(x)] - \frac{1}{2}\right| \leq \frac{1}{n^{\Omega(1)}}$.

Using Lemma 5.0.4, it follows that f is an extractor for $(n^{1-\delta}, t, \gamma)$-non-oblivious bit-fixing sources with error $1/n^{\Omega(1)}$. □

Proof of Theorem 14. Let reduce : $\{0,1\}^n \times \{0,1\}^n \to \{0,1\}^D$ be the function from Theorem 6.4.1 with $t = c(\log n)^{18}$, where c is the constant from Theorem 13. Set the constant $C = 74$ and $C_1 = c'$, where c' is the constant from Theorem 6.4.1. We note that $D = n^{O(1)}$.

Let bitExt : $\{0,1\}^D \to \{0,1\}$ be the explicit extractor from Theorem 13 set to extract from (q, t, γ)-non-oblivious bit-fixing source on D bits with error $\frac{1}{n^{\Omega(1)}}$, where $q = D^{1-\delta}$ and $\gamma \leq 1/D^{t+1}$.

Define

$$2\text{Ext}(x, y) = \text{bitExt}(\text{reduce}(x, y)).$$

Let $\mathbf{X}$ and $\mathbf{Y}$ be any two independent (n, k)-sources, where $k \geq C_1(\log n)^C$. We prove that

$$|(2\text{Ext}(\mathbf{X}, \mathbf{Y}), \mathbf{Y}) - (\mathbf{U}_1, \mathbf{Y})| \leq \frac{1}{n^{\Omega(1)}}.$$

Let $\mathbf{Z} = \text{reduce}(\mathbf{X}, \mathbf{Y})$. Theorem 6.4.1 implies that with probability at least $1 - n^{-\omega(1)}$ (over $y \sim \mathbf{Y}$), the conditional distribution $\mathbf{Z}|\mathbf{Y} = y$ is a (q, t, γ)-non-oblivious bit-fixing source on M bits. Thus, for each such y,

$$|\text{bitExt}(\text{reduce}(\mathbf{X}, y)) - \mathbf{U}_1| \leq \frac{1}{n^{\Omega(1)}}.$$

Thus, we have

$$|(2\text{Ext}(\mathbf{X}, \mathbf{Y}), \mathbf{Y}) - (\mathbf{U}_1, \mathbf{Y})| \leq \frac{1}{n^{\omega(1)}} + \frac{1}{n^{\Omega(1)}}.$$

□

6.6 Achieving Smaller Error

We show that it is indeed possible to achieve an extractor with smaller error at the expense of increasing the running time of the extractor. We achieve this by slightly modifying the construction in Theorem 14.

Informally, we now use the sources $\mathbf{X}$ and $\mathbf{Y}$ to generate a much longer string $\mathbf{Z}$ with the property that most of the bits are t-wise independent. This allows us to achieve smaller error in the reduction, and now applying the extractor for (q, t, γ)-sources developed in Theorem 13, the result follows.

Theorem 6.6.1 (Theorem 15 restated)**.** *There exists a constant $C > 0$ such that for all $n \in \mathbb{N}$ and any $\epsilon > 0$, there exists a 2-source extractor $2\text{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ computable in time $\text{poly}(n, 1/\epsilon)$ for min-entropy at least $\log^C(n/\epsilon)$ and error ϵ.*

Proof sketch. We provide the details of the construction and omit the proof since it is very similar to the proof of Theorem 14.

We set up the required ingredients as follows:

- Let $t = b(\log(5D/\epsilon))^{18}$, where b is the constant from Theorem 5.2.1.
- Let $\text{nmExt} : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}$ be a (t, k, ϵ_1)-non-malleable extractor from Theorem 1. Thus $d_1 = c_1 t^2 \log^2(n/\epsilon_1)$, for some constant c_1. For such an extractor to exists, we require $k \geq \lambda_1 t \log^2(n/\epsilon_1)$.
- Let $\text{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^{d_1}$ be the seeded extractor from Theorem 2.1.3 set to extract from min-entropy $k/2$ with error ϵ_2. Thus, $d = c_2 \log(n/\epsilon_2)$, for some constant c_2. Let $D = 2^d = (n/\epsilon_2)^{c_2}$. Such an extractor exists for $k \geq 3d_1$.
- Choose $\delta > 0$, such that $\delta' = \delta c_2 < 9/10$.
- Let $f : \{0,1\}^D \to \{0,1\}$ be the function from Theorem 5.3.1 such that f is $\mathbf{I}_q(f) \leq q/D^{1-\frac{\delta}{2}}$ and $\left|\mathbf{E}_{v \sim \mathbf{U}_D}[f(v)] - \frac{1}{2}\right| \leq D^{-\beta}$ for some small constant β.
- Pick ϵ_1, ϵ_2 such that the following inequalities are satisfied:
  - $D = (n/\epsilon_2)^{c_2} \geq \max\{1/\epsilon^{1/\beta}, 1/\epsilon^{2/\delta}\}$,
  - $\epsilon_2 \leq D^{-\delta}/2 = (\epsilon_2/n)^{\delta'}$,

– $\sqrt{\epsilon_1} \leq \frac{1}{5tD^{t+1}}$.

Thus, we can pick $\epsilon_2 = \min\{n\epsilon^{\frac{1}{c_2\beta}}, n\epsilon^{\frac{2}{c_2\delta}}, 1/n^{\delta'/(1-\delta')}\}$ and $\epsilon_1 = 1/(5tD^{t+1})$.

- With this setting of parameters, we require $k \geq (\log(n/\epsilon))^{c'}$, where c' is a large enough constant, for nmExt and Ext to work.

Let $\{0,1\}^{d_2} = \{r_1, \ldots, r_{D_2}\}$. Define

$$\text{reduce}(x, y) = \text{nmExt}(x, \text{Ext}(y, r_1)) \circ \ldots \circ \text{nmExt}(x, \text{Ext}(y, r_{D_2}))$$

and

$$2\text{Ext}(x, y) = f(\text{reduce}(x, y)).$$

Using arguments similar to the proof of Theorem 14, it can be shown that 2Ext is an extractor for min-entropy k with error $O(\epsilon)$. Further, the extractor runs in time $\text{poly}(n, 1/\epsilon)$. □

6.7 Towards Optimal Ramsey Graphs

Since one of the motivations to study 2-source extractors is the connection to Ramsey graphs and to meet Erdős' challenge to explicitly construct $O(\log N)$-Ramsey graphs on N vertices, it is interesting to see if the above framework can be pushed to meet this goal. After our work in [CZ16a], Meka [Mek15] improved one of the components in the above construction. In joint work with Xin Li [CL16a], we construct an improved t-non-malleable extractor (see Theorem 11, Chapter 4). Using these components in the framework developed, the following results are easy to obtain by suitably optimizing parameters.

Theorem 17. *There exists a constant $C > 0$ such that for any $\delta > 0$ and for all $n, k \in \mathbb{N}$ with $k \geq C(\log n)^{2\sqrt{6(1+\delta)}+3}$ and any constant $\epsilon < \frac{1}{2}$, there exists an efficient polynomial time computable 2-source extractor min-entropy k with error ϵ that outputs 1 bit.*

Theorem 18. *There exists a constant $C > 0$ such that for any $\delta > 0$ and for all $n, k \in \mathbb{N}$ with $k \geq C(\log(n))^{4\sqrt{5(1+\delta)}+5}$, there exists an efficient polynomial time computable 2-source extractor min-entropy k with error $n^{-\Omega(1)}$ and output length $\Omega(k)$.*

Thus Theorem 17 implies K-Ramsey graphs on $N = 2^n$ vertices, with $K = 2^{(\log\log n)^{7.899}}$. This currently stands as the best known explicit construction of a Ramsey graph.

Chapter 7

Multi-Source Extractors

[1] In Chapter 6 we studied the problem of extracting from 2 independent sources. As we saw, the best known 2-source extractor requires min-entropy roughy $\log^8 n$ (for constant error). Recall that any explicit (k, ϵ)-2-source extractor $\text{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ (for any $\epsilon < 1/2$) implies a 2^k-Ramsey graph on $N = 2^n$ vertices. Thus achieving min-entropy $k = \log n + O(1)$ immediately implies an explicit $O(\log N)$-Ramsey graph matching Erdős' challenge from the 1940's. However based on the current methods to construct 2-source extractors (see Chapter 6), it looks like a challenging task to even achieve min-entropy $O(\log n)$.

In this chapter we study a relaxed version of the problem, and allow the extractor access to multiple independent source. We formally define a multi-source extractor.

Definition 7.0.1. *A function* $\text{iExt} : (\{0,1\}^n)^C \to \{0,1\}^m$ *is an extractor for* C *independent sources with min-entropy* k *and error* ϵ *if for any independent* (n,k)*-sources* $\mathbf{X}_1, \ldots, \mathbf{X}_C$*, we have*

$$|\text{iExt}(\mathbf{X}_1, \ldots, \mathbf{X}_C) = \mathbf{U}_m| \leq \epsilon.$$

An impressive line of work studied this problem and constructed extractors with excellent parameters [BIW06, BKS+10, Rao09a, BRSW12, RZ08, Li11a, Li13b, Li13a, Li15e, Coh15a]. However,

[1] parts of this chapter have been previously published [CL16a]

the smallest entropy these constructions can achieve is $(\log n)^{2+\delta}$ for any constant $\delta > 0$ [Li13a], which uses $O(1/\delta) + O(1)$ sources. In a very recent work, Cohen and Schulman [CS16] managed to break this "quadratic" barrier, and constructed extractors for $O(1/\delta) + O(1)$ sources, each having min-entropy at least $(\log n)^{1+\delta}$.

The results in this chapter are based on joint work with Xin Li [CL16a].

7.1 Our Result and Overview of techniques

Our main result in this section is the following.

Theorem 7.1.1. *There exists a constant $C > 0$ s.t for all $n, k \in \mathbb{N}$ and any constant $\epsilon > 0$, with $k \geq 2^{C\sqrt{\log\log(n)}} \log n$, there exists an explicit function* $\mathrm{Ext} : (\{0,1\}^n)^C \to \{0,1\}$*, such that*

$$|\mathrm{Ext}(\mathbf{X}_1, \ldots, \mathbf{X}_C) - \mathbf{U}_1| \leq \epsilon.$$

On a high level, we follow a framework introduced by Cohen and Schulman [CS16], and improve a key component of their construction which allows us to achieve the improved result. The first step in [CS16] is to use $O(1)$ (an absolute constant) independent sources and transform it into a collection of r matrices such that at least $r - r^{0.49}$ of these matrices are 'good' and follow a certain independence property. In particular, for any good matrix $\mathbf{X}$ and any distinct t of the other good matrices $\mathbf{X}^1, \ldots, \mathbf{X}^t$, there exists a row index h such that $(\mathbf{X}_h, \mathbf{X}^1_h, \ldots, \mathbf{X}^t_h) \approx (\mathbf{U}_m, \mathbf{X}^1_h, \ldots, \mathbf{X}^t_h)$, where $t = O(1)$ is some parameter. The next idea is to use an independence preserving merger (IPM), which by definition, uses a few additional sources and transforms these matrices into a r.v $\mathbf{Z}$ on r bits such that at least $r - r^{0.49}$ bits of $\mathbf{Z}$ are almost t-wise independence. By using our explicit non-malleable independence preserving merger construction (NIPM) from Chapter 3, we show how to construct an improved IPM which uses just 1 additional source (this is the step where [CS16] uses an additional $O(1/\delta)$ sources). It is known that the majority function [DGJ+10, Vio14] is an extractor for $\mathbf{Z}$ (see Lemma 7.3.3), which completes the construction. Therefore, we obtain

a multi-source extractor for an absolute constant number of $(n, \log^{1+o(1)} n)$-sources, which outputs one bit with constant (or slightly sub-constant) error.

We first present our IPM construction in the next section, and use this to improve upon the results on multi-source extractors obtained in [CS16] in Section 7.3.

7.2 An Independence Preserving Merger Using a Weak Source

An important ingredient in our construction is an explicit construction of an independence preserving merger. We use the (L, ℓ, t)-NIPM constructed in the Section 4.4.1 to merge the r.v's $\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t$, each supported on boolean $L \times m$ matrices, with the guarantee that there is some $h \in [L]$ s.t $\mathbf{X}_h$ is uniform on average conditioned on $\{\mathbf{X}^g_h : g \in [t]\}$ using an independent (n, k)-source $\mathbf{Y}$ (instead of a seed as in the previous section). Our construction improves the construction of an IPM by Cohen and Schulman [CS16], and further uses just 1 independent source.

Recall that for any $a \times b$ matrix $\mathbf{V}$, and any $S \subseteq [a]$, we use $\mathbf{V}_S$ to denote the matrix obtained by restricting $\mathbf{V}$ to the rows indexed by S.

Our main result in this section is the following theorem.

Theorem 7.2.1. *For all integers $m, \ell, L, t > 0$, any $\epsilon > 0$, $r = \lceil \frac{\log L}{\log \ell} \rceil$ and any $k \geq 2c_{3.5.7}\ell \log(m/\epsilon)(t+2)^{r+2}$, there exists an explicit function (L, ℓ, t)-IPM $: \{0,1\}^{mL} \times \{0,1\}^n \to \{0,1\}^{m''}$, $m'' = (0.9/t)^{r+1}(m - c_{3.5.7}\ell(t+1)r\log(m/\epsilon) - c_{2.1.2}(t+2)\log(n/\epsilon)))$, such that if the following conditions hold:*

- *$\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t$ are r.v's, each supported on boolean $L \times m$ matrices s.t for any $i \in [L]$, $|\mathbf{X}_i - \mathbf{U}_m| \leq \epsilon$,*
- *$\mathbf{Y}$ is an (n, k)-source, independent of $\{\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t\}$.*
- *there exists an $h \in [\ell]$ such that $|(\mathbf{X}_h, \mathbf{X}^1_h, \ldots, \mathbf{X}^t_h) - (\mathbf{U}_m, \mathbf{X}^1_h, \ldots, \mathbf{X}^t_h)| \leq \epsilon$,*

then

$$|(L,\ell,t)\text{-IPM}(\mathbf{X},\mathbf{Y}),(L,\ell,t)\text{-IPM}(\mathbf{X}^1,\mathbf{Y}),\ldots,(L,\ell,t)\text{-NIPM}(\mathbf{X}^t,\mathbf{Y})$$
$$-\mathbf{U}_{m''},(L,\ell,t)\text{-IPM}(\mathbf{X}^1,\mathbf{Y}),\ldots,(L,\ell,t)\text{-IPM}(\mathbf{X}^t,\mathbf{Y})| \leq 3c'_{3.5.7}L\epsilon.$$

Proof. We set up parameters and ingredients required in our construction.

- Let $d = 0.8k, d' = c_{2.1.2}\log(m/\epsilon), d_1 = c_{2.1.2}\log(n/\epsilon)$.
- Let $\text{Ext}_1 : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^d$ be a (k,ϵ)-strong-seeded extractor from Theorem 2.1.2.
- Let $\text{Ext}_2 : \{0,1\}^m \times \{0,1\}^{d'} \to \{0,1\}^{m'}$, $m' = 0.9(m - c_{2.1.2}(t+1)\log(n/\epsilon))$, be a $(m - c_{2.1.2}(t+1)\log(n/\epsilon),\epsilon)$-strong-seeded extractor from Theorem 2.1.2.
- Let (L,ℓ,t)-NIPM : $\{0,1\}^{Lm'} \times \{0,1\}^d \to \{0,1\}^{m''}$ be the function from Theorem 4.4.4 with error parameter ϵ.

Algorithm 8: (L,ℓ,t)-IPM(x,y)
Input: x is a boolean $L \times m$ matrix, and y is a bit string of length n.
Output: A bit string of length m''.

1 Let $w = \text{Slice}(x_1, d_1)$
2 Let $z = \text{Ext}_1(y,w)$.
3 Let $v = \text{Slice}(z,d')$.
4 Let $\overline{v}$ be a $L \times m'$-matrix, whose i'th row is given by $\overline{v}_i = \text{Ext}_2(x_i,v)$.
5 Output $\overline{z} = (L,\ell,t)$-NIPM$(\overline{v},z)$.

We begin by proving the following claim.

Claim 7.2.2. *Conditioned on* $\mathbf{W}, \{\mathbf{W}^g : g \in [t]\}$*, the following hold:*

- $\mathbf{Z}$ *is* ϵ*-close to* $\mathbf{U}_d$*,*
- $\mathbf{Z}, \{\mathbf{Z}^g : g \in [t]\}$ *is independent of* $\mathbf{X}, \{\mathbf{X}^g : g \in [t]\}$*,*
- *For each* $i \in [L]$*,* $\mathbf{X}_i$ *has average conditional min-entropy at least* $m - (t+2)\log(n/\epsilon)$*,*

- $\mathbf{X}_h|\{\mathbf{X}_h^g : g \in [t]\}$ *has average conditional min-entropy at least* $m - (t+2)d_1 \log(n/\epsilon)$.

Proof. Since Ext_1 is a strong extractor, we can fix $\mathbf{W}$, and $\mathbf{Z}$ is ϵ-close to $\mathbf{U}_d$ on average. Further, $\mathbf{Z}$ is now a deterministic function of $\mathbf{X}_1$. Thus, we can fix $\{\mathbf{W}^1, \ldots, \mathbf{W}^t\}$, without affecting the distribution of $\mathbf{Z}$. Since $\mathbf{W}^i$ is on d_1 bits, and without any prior conditioning since $\mathbf{X}|\{\mathbf{X}_h^g : g \in [t]\}$ is ϵ-close to uniform on average, it follows that conditioned on $\{\mathbf{X}_h^g : g \in [t]\}, \mathbf{W}, \{\mathbf{W}^g : g \in [t]\}$, the r.v $\mathbf{X}_h$ has average conditional min-entropy $m - (t+1)d_1 \log(n/\epsilon) - \log(1/\epsilon)$. □

Claim 7.2.3. *Conditioned on* $\mathbf{W}, \{\mathbf{W}^g : g \in [t]\}, \mathbf{V}, \{\mathbf{V}^g : g \in [t]\}$, *the following hold:*

- $\{\mathbf{Z}, \mathbf{Z}^1, \ldots, \mathbf{Z}^t\}$ *is independent of* $\{\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t\}$,
- $\{\overline{\mathbf{V}}, \overline{\mathbf{V}}^1, \ldots, \overline{\mathbf{V}}^t\}$ *is a deterministic function of* $\{\mathbf{X}, \mathbf{X}^1, \ldots, \mathbf{X}^t\}$,
- *For each* $i \in [L]$, $\overline{\mathbf{V}}_i$ *is* 2ϵ*-close to uniform,*
- $\overline{\mathbf{V}}_h|\{\overline{\mathbf{V}}_h^g : g \in [t]\}$ *is* 2ϵ*-close to uniform on average.*
- $\mathbf{Z}$ *has average conditional min-entropy at least* $d - (t+2)\log(m/\epsilon)$.

Proof. Fix $\mathbf{W}, \{\mathbf{W}^g : g \in [t]\}$. Thus, by Claim 7.2.2, we have

- $\mathbf{Z}$ is ϵ-close to $\mathbf{U}_d$,
- $\mathbf{Z}, \{\mathbf{Z}^g : g \in [t]\}$ is independent of $\mathbf{X}, \{\mathbf{X}^g : g \in [t]\}$,
- For each $i \in [L]$, $\mathbf{X}_i$ has average conditional min-entropy at least $m - (t+2)\log(n/\epsilon)$,
- $\mathbf{X}_h|\{\mathbf{X}_h^g : g \in [t]\}$ has average conditional min-entropy at least $m - (t+2)\log(n/\epsilon)$.

Since each $\mathbf{X}_i$ has average conditional min-entropy at least $m - (t+2)\log(n/\epsilon)$, it follows that each $\overline{\mathbf{V}}_i$ is 2ϵ-close to uniform and Ext_2 is a strong extractor, it follows that $\overline{\mathbf{V}}_i$ is 2ϵ-close to $\mathbf{U}_d$ on average even conditioned on $\{\mathbf{V}, \mathbf{V}^1, \ldots, \mathbf{V}^t\}$. After this fixing, $\mathbf{Z}$ has average conditional min-entropy at least $d - (t+2)\log(n/\epsilon)$.

We now prove that $\overline{\mathbf{V}}_h|\{\overline{\mathbf{V}}_h^g : g \in [t]\}$ is 2ϵ-close to uniform on average. First, we fix the r.v's $\mathbf{W}, \{\mathbf{W}^g : g \in [t]\}$ (at this point no other r.v's are fixed). As before, we have $\mathbf{X}_h|\{\mathbf{X}_h^g : g \in [t]\}$ has average conditional min-entropy $k_x \geq m - (t+2)\log(n/\epsilon)$. Thus, we fix $\{\mathbf{X}_h^g : g \in [t]\}$. Now since Ext_2 is a strong extractor, $\overline{\mathbf{V}}_h$ is uniform on average even conditioned on $\mathbf{V}$. We fix $\mathbf{V}$, and thus $\overline{\mathbf{V}}_h$ is a deterministic function of $\mathbf{X}_h$. Further, $\{\overline{\mathbf{V}}_h^g : g \in [t]\}$ is a deterministic function of $\{\mathbf{V}^g : g \in [t]\}$, and hence a deterministic function of $\mathbf{Z}, \{\mathbf{Z}^g : g \in [t]\}$. Thus, we can fix $\{\overline{\mathbf{V}}_h^g : g \in [t]\}$ without affecting the distribution of $\overline{\mathbf{V}}_h$. This completes the proof of our claim. □

The correctness of the function IPM is direct from the next claim.

Claim 7.2.4. *Conditioned on $\{\overline{\mathbf{Z}}^g : g \in [t]\}\}$, the r.v $\overline{\mathbf{Z}}$ is $3L\epsilon$-close to uniform on average.*

Proof. Fix the r.v's $\mathbf{W}, \{\mathbf{W}^g : g \in [t]\}, \mathbf{V}, \{\mathbf{V}^g : g \in [y]\}$. We observe that the following hold:

- $\mathbf{Z}, \{\mathbf{Z}^g : g \in [t]\}$ is independent of $\mathbf{Y}, \{\mathbf{Y}^g : g \in [t]\}$,
- For each $i \in [L]$, $\overline{\mathbf{V}}_i$ is 2ϵ-close to uniform,
- $\overline{\mathbf{V}}_h|\{\overline{\mathbf{V}}_h^g : g \in [t]\}$ is 2ϵ-close to uniform on average.
- $\mathbf{Z}$ has average conditional min-entropy at least $d - (t+2)\log(m/\epsilon)$.

The claim is now direct from Theorem 4.4.4 by observing that by our choice of parameters, the following hold:

- $d \geq (c_{3.5.7}\ell \log(m/\epsilon) + d'')(t+2)^{r+1}$, where $d'' = (t+2)\log(m/\epsilon)$,
- $\mathbf{Z}$ has average conditional min-entropy at least $d - d''$,
- $m'' \leq (0.9/t)^r(m' - c_{3.5.7}\ell(t+1)r\log(m/\epsilon))$.

This completes the proof of the claim, and hence Theorem 7.2.1 follows. □

□

7.3 The Extractor Construction

We recall a reduction by Cohen and Schulman [CS16]. Informally, they used a constant number of independent sources to transform into a sequence of matrices such that a large fraction of these matrices follow a certain t-wise independence property. For our purposes, we need to slightly modify this construction. The length of the rows (the parameter m in the following theorem) in the work of [CS16] can be set to $c\log(n/\epsilon)$, for any constant c. Using another additional source and extracting from it using each row as seed (using any optimal strong-seeded extractor), the length of each row can be made $\Omega(k)$.

We state the theorem from [CS16] with this modification.

Theorem 7.3.1 ([CS16])**.** *There exist constants $\alpha > 0$ and and $c_{7.3.1}$ such that for all $n, t \in \mathbb{N}$, and for any $\epsilon, \delta > 0$, there exists an polynomial time computable function $f : (\{0,1\}^n)^C \to (\{0,1\}^{Lm})^r$, where $C = 7/\alpha, L = O(t \log n), r = n^{3/\alpha}, m = \Omega(k)$, such that the following hold: Let $\mathbf{X}_1, \ldots, \mathbf{X}_C$ be independent (n,k) sources, $k = c_{7.3.1} t \log(t) \log(n \log t/\epsilon)$. Then there exists a subset $S \subset [r]$, $|S| \geq r - r^{\frac{1}{2}-\alpha}$ and a sequence of $L \times m$ matrices $\mathbf{Y}^1, \ldots, \mathbf{Y}^r$ such that:*

- *$f(\mathbf{X}_1, \ldots, \mathbf{X}_C)$ is $1/r$-close to $\mathbf{Y}^1, \ldots, \mathbf{Y}^r$,*
- *for any $i \in [L]$ and $g \in S$, $\mathbf{Y}^g_i$ is ϵ-close to $\mathbf{U}_m$,*
- *for any $g \in S$, and any distinct $i_1, \ldots, i_t$ in $S \setminus \{g\}$, there exists an $h \in [L]$ such that $\mathbf{Y}^g_h | \{\mathbf{Y}^j_h : j \in [r] \setminus \{g\}\}$ is ϵ-close to uniform.*

We are now ready to present our extractor construction. By composing Theorem 7.3.1 with our independence preserving merger from Section 7.2, we have the following result.

Theorem 7.3.2. *There exists a constant $\alpha > 0$ such that for all $n, t \in \mathbb{N}$, and for any $\epsilon, \delta > 0$, there exists an polynomial time computable function* $\mathrm{reduce} : (\{0,1\}^n)^{C+1} \to \{0,1\}^r$*, where $C = \frac{7}{\alpha} + 1, r = n^{3/\alpha}$, such that the following hold: Let $\mathbf{X}_1, \ldots, \mathbf{X}_C$ be independent (n,k) sources, $k \geq 2^{\sqrt{\log t + \log\log n}} \log(k/\epsilon)(t+2)^{O(\sqrt{\log t + \log\log n})} + c_{7.3.1} t \log(t) \log(n \log t/\epsilon)$, and let $\mathbf{Z} = \mathrm{reduce}(\mathbf{X}_1, \ldots, \mathbf{X}_{C+1})$.*

Then there exists a subset $S \subset [r]$, $|S| \geq r - r^{\frac{1}{2}-\alpha}$ such that $\mathbf{Z}_S$ is $n^{-\Omega(1)}$-close to a $(t, \gamma_{7.3.2})$-wise independent distribution, where $\gamma_{7.3.2} = O(\epsilon t \log n)$.

Proof. Let $f : (\{0,1\}^n)^C \to (\{0,1\}^{Lm})^r$ be the function from Theorem 7.3.1 with $\epsilon_{7.3.1} = \epsilon, m = \beta k$ for some constant $\beta > 0$. Thus $L = O(t \log n)$. Let (L, ℓ, t)-IPM : $(\{0,1\}^{Lm})^t \times \{0,1\}$ be the function from Theorem 7.2.1, with $\ell = 2^{\sqrt{\log L}} = 2^{O(\sqrt{\log t + \log\log n})}$ and error parameter $\epsilon_{7.2.1} = \epsilon$. Define

$$\text{reduce}(x_1, \ldots, x_{C+1}) = (L, \ell, t)\text{-IPM}(f(x_1, \ldots, x_C), x_{C+1}).$$

We note that $k > c_{7.3.1} t \log(t) \log(n \log t/\epsilon)$. Thus, using Theorem 7.3.1, it follows that there exists a subset $S \subset [r]$, $|S| \geq r - r^{\frac{1}{2}-\alpha}$ and a sequence of $L \times m$ matrices $\mathbf{Y}^1, \ldots, \mathbf{Y}^r$ such that:

- $f(\mathbf{X}_1, \ldots, \mathbf{X}_C)$ is $1/r$-close to $\mathbf{Y}^1, \ldots, \mathbf{Y}^r$,
- for any $i \in [L]$ and $g \in S$, $\mathbf{Y}^g_i$ is ϵ-close to $\mathbf{U}_m$,
- for any $g \in S$, and any distinct $i_1, \ldots, i_t$ in $S \setminus \{g\}$, there exists an $h \in [L]$ such that $\mathbf{Y}^g_h | \{\mathbf{Y}^j_h : j \in [r] \setminus \{g\}\}$ is ϵ-close to uniform.

We now work with the sources $\mathbf{Y}^1, \ldots, \mathbf{Y}^r$, and add an error of $1/r$ in the end. The theorem is now direct using Theorem 7.2.1 and observing that the following hold by our setting of parameters:

- $k \geq 2c_{3.5.7} \ell \log(k/\epsilon)(t+2)^{\lceil \frac{\log L}{\log \ell} \rceil + 1}$,
- $m = \beta k \geq 2^{\sqrt{\log L}}(c_{3.5.7} \ell (t+1) r \log(m/\epsilon) + c_{2.1.2}(t+2)\log(n/\epsilon))$.

□

Our multi-source extractor in Theorem 7.1.1 is now easy to obtain using a result on the majority function.

Theorem 7.3.3 ([DGJ+10,Vio14,CS16])**.** *Let $\mathbf{Z}$ be a source on r bits such that there exists a subset $S \subset [r]$, $|S| \geq r - r^{\frac{1}{2}-\alpha}$ such that $\mathbf{Z}_S$ is t-wise independent. Then,*

$$\left|\Pr[\text{Majority}(\mathbf{Z}) = 1] - \frac{1}{2}\right| \leq O\left(\frac{\log t}{t} + r^{-\alpha}\right).$$

We also recall a result about almost t-wise independent distributions.

Theorem 7.3.4 ([AGM03])**.** *Let $\mathcal{D}$ be a (t,γ)-wise independent distribution on $\{0,1\}^n$. Then there exists a t-wise independent distribution that is $n^t\gamma$-close to $\mathcal{D}$.*

Thus, we have the following corollary.

Corollary 7.3.5. *There exists a constant c such that the following holds: Let $\mathbf{Z}$ be a source on r bits such that there exists a subset $S \subset [r]$, $|S| \geq r - r^{\frac{1}{2}-\alpha}$ such that $\mathbf{Z}_S$ is (t,γ)-wise independent. Then,*

$$|\Pr[\text{Majority}(\mathbf{Z}) = 1] - \frac{1}{2}| \leq c\left(\frac{\log t}{t} + r^{-\alpha} + \gamma r^t\right).$$

Proof of Theorem 7.1.1. Set t to a large enough constant such that $\frac{c\log t}{t} < \epsilon/2$. Let α be the constant from Theorem 7.3.2, $r = n^{3/\alpha}$ and $C = \frac{7}{\alpha} + 1$. Let reduce be the function from Theorem 7.3.2 with parameter $t_{7.3.2} = t$, $r_{7.3.2} = r$, and the error parameter $\epsilon_{7.3.2}$ set such that the parameter $\gamma_{7.3.2} \leq \frac{1}{r^{t+1}}$. This can be ensured by setting $\epsilon = n^{-C'}$ for a large enough constant C'.

Define

$$\text{Ext}(x_1, \ldots, x_C) = \text{Majority}(f(x_1, \ldots, x_C)).$$

Let $\mathbf{Z} = f(\mathbf{X}_1, \ldots, \mathbf{X}_C)$. We note that with this setting of parameters, there exists some constant C'' such that any $k \geq 2^{C''\sqrt{\log\log n}}\log(n)$ is sufficient for the conclusion of Theorem 7.3.2 to hold. Thus, $\mathbf{Z}$ is a source on r bits such that there exists a subset $S \subset [r]$, $|S| \geq r - r^{\frac{1}{2}-\alpha}$ for which $\mathbf{Z}_S$ is (t,γ)-wise independent. Theorem 7.1.1 is now direct from Corollary 7.3.5. □

Chapter 8

Extractors for Sumset Sources

[1] This chapter is based on , we introduce and study a new model of weak sources which we call *sumset sources*. Informally, this is the class of sources which are the sum (XOR) of independent sources. This further reduces the assumptions made on weak sources, and provides a unified framework for designing extractors for many well studied classes of sources. We then construct explicit extractors for sumset sources and apply them to other classes of sources studied before. In several cases we obtain substantial improvements over previous constructions. We now formally define sumset sources.

Definition 8.0.1. *For any two strings $x, y \in \{0,1\}^n$, define $x + y$ to be the bit wise* XOR *of the two strings.*

Definition 8.0.2 ((n,k,C)-sumset source)**.** *A weak source $\mathbf{X}$ is called an (n,k,C)-sumset source if $\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_C$, where $\mathbf{X}_1, \ldots, \mathbf{X}_C$ are independent (n,k)-sources.*

A well known extractor for this class of sources is based on the Paley graph function (see Theorem 2.5.4) and works for the sum of 2 independent sources, with one source having min-entropy at least $> n/2$ and the other having min-entropy $> \log n$. On the other extreme, the work of Kamp et al. [KRVZ11] shows how to extract when $\mathbf{X}$ is a sum of exponentially many sources when the

[1] parts of this chapter have been previously published [CL16b]

sum of the min-entropies of these sources is large enough. To the best of our knowledge, there is no other known explicit construction for $2 \leq C \leq 2^{O(n)}$.

The results in this chapter are based on joint work with Xin Li [CL16b].

Our main result is an explicit construction of an extractor for the sum of a constant number of independent sources, each containing polylogarithmic min-entropy.

Theorem 19. *There exist constants $c, C > 0$ and a small constant $\beta > 0$ such that for all $n \in \mathbb{N}$ and $k \geq \log^c n$, there exists a polynomial time computable extractor for (n, k, C)-sumset sources, with error $n^{-\Omega(1)}$ and output length k^β.*

8.1 Relations and Applications to Other Sources

Independent Sources

The class of independent sources is clearly a special case of sumset sources. That is, if we view the joint distribution of several independent sources as one source $\mathbf{X}$, then $\mathbf{X}$ is also a sumset source. Thus, our construction in Theorem 19 also gives an extractor for a constant number of independent sources with polylogarithmic min-entropy. If we can improve the construction and obtain an explicit extractor for $(n, k, 2)$-sumset sources with $k \geq \log^c n$, then this will also match the two source extractors in [CZ16a, Li15b].

Affine Sources

An affine source X on n bits with entropy k is the uniform distribution over some unknown affine subspace of dimension k in $\{0,1\}^n$ (viewing $\{0,1\}^n$ as $\mathbb{F}_2^n$[2]). This model generalizes oblivious bit-fixing sources (where some of the bits are uniform and independent, while others are fixed) and thus has received attention for its applications to cryptography. Affine extractors have also been used by Viola [Vio14] to construct extractors for sources generated by NC^0 and AC^0 circuits. Further,

[2] In general, affine sources can be defined on any field $\mathbb{F}_q$, but in this paper we focus on $\mathbb{F}_2$.

good affine extractors imply the best known circuit lower bounds [DK11, FGHK15].

Using the probabilistic method, one can show that affine extractors exist for entropy $k = O(\log n)$. However until recently, the best known explicit constructions for affine extractor was due to Bourgain [Bou07], who using sophisticated techniques from additive combinatorics and gave an explicit extractor for min-entropy at least δn, for any constant δ. This construction was subsequently slightly improved to entropy $n/\sqrt{\log \log n}$ by Yehudayoff [Yeh11] and Li [Li11b]. In a very recent work, Li [Li15c] constructed the first explicit affine extractors for polylogarithmic entropy.

We note that an affine source is also a special case of sumset source, since an affine subspace of dimension k can be written as the sum of C affine subspaces of dimension k/C. Thus, as a direct corollary of our extractor for sumset sources, we also obtain extractors for affine sources with polylogarithmic min-entropy, matching the recent work of Li [Li15c].[3]

Corollary 8.1.1. *There exists a constant $c > 0$ and a small constant $\beta > 0$ such that for all $n, k \in \mathbb{N}$ with $k \geq \log^c n$, there exists a polynomial time computable extractor for affine sources in $\{0,1\}^n$ with entropy k. The extractor has error $n^{-\Omega(1)}$ and output length k^β.*

Proof. Let $\mathbf{X}$ be an affine source with min-entropy k. Let $v_1, \ldots, v_k$ be a basis of $\mathbf{X}$ and b be the shift vector. Let C be the constant in Theorem 19. For $i \in [C]$, define the source $\mathbf{X}_i$ to be the uniform distribution on the linear subspace spanned by $v_{(i-1)k/C+1} \ldots, v_{ik/C}$ for $i = 2, \cdots, C$, and define $\mathbf{X}_1$ to be the uniform distribution on the affine subspace spanned by $v_1 \ldots, v_{k/C}$ with shift vector b. Thus $\mathbf{X} = \sum_{j=1}^{C} \mathbf{X}_i$, where each $\mathbf{X}_i$ has min-entropy k/C and the $\mathbf{X}_i$'s are independent. Thus $\mathbf{X}$ is a $(n, k/C, C)$-sumset source, and we can now apply Theorem 19. □

Small-Space Sources

We study small-space sources in Chapter 9 and refer the reader to this chapter for more details.

[3] The extractor construction is essentially the same as in [Li15c], but the analysis is different.

Interleaved Sources

We study interleaved sources in Chapter 10 and refer the reader to this chapter for more details.

Total Entropy Independent Sources and Somewhere Entropy Independent Sources

We study these sources in Chapter 9 and refer the reader to this chapter for more details.

8.2 Overview of Techniques

On a very high level, our extractor follows the same spirit of our 2-source extractor construction in Chapter 6. That is, we first convert our sumset source into a (N^δ, t, γ)-NOBF source on N bits (see Chapter 5 for a definition of NOBF sources) , where $N = n^{O(1)}, t = k^\alpha, \gamma < 1/N^{t+1}$, for some constants $0 < \delta, \alpha < 1$. We will then apply extractors for this source constructed in Chapter 5.

To obtain such a non-oblivious bit-fixing source, it suffices to use two *independent* sources as shown in Chapter 6. More specifically, if we have a somewhere random source[4] with N rows such that $N - N^\delta$ rows are uniform, then it is not hard to show that we can use an explicit correlation breaker from Chapter 3 (Theorem 3.4.2), we obtain an NOBF source with at least $N - N^\delta$ 'good' bits that are k^α-wise independent.

Now the problem is how to obtain the somewhere random source. The standard way is to use a seeded extractor with seed length $O(\log n)$ (so that to keep the running time polynomial in n) and try all possible values of the seed. Each seed will give an output and we can then concatenate the output to form a matrix. This does indeed give us a somewhere random source, however there are now two problems. First, we cannot just use any seeded extractor with seed length $O(\log n)$. This is because we need to apply the seeded extractor to the sum of several independent sources, and we need to keep the "sum" structure carefully for the purpose of alternating extraction later. If we just use any seeded extractor, then after applying the extractor the "sum" structure may not

[4] A somewhere random source is a matrix of random variables such that at least one row is uniform.

be preserved. Therefore, here again we need to use a linear seeded extractor. Luckily, we do have linear seeded extractors with seed length $O(\log n)$, due to a construction in [Li15c].

Second, just doing this is not enough, since the error of the somewhere random source is not good enough for our purpose. Specifically, in order to apply the extractor for non-oblivious bit-fixing source we need the error to be negligibly small, while the error we obtained from a seeded extractor with seed length $O(\log n)$ is only polynomially small. Note this is different from the affine extractor construction in [Li15c], as in the case of affine sources one can show that if we use a linear seeded extractor, then most of the rows in the somewhere random source actually have error 0. However for general weak random sources the best error one can hope for (even with a linear seeded extractor) is $1/\text{poly}(n)$ if the seed length is $O(\log n)$.

To get around this, we use a sampling method (similar to a technique seen in Chapter 6). Specifically, they first used an extractor (or a non-malleable extractor) with large seed length to achieve small error from one source, and then use another independent source to sample from the rows of the somewhere random source to bring down the number of rows. The first idea would be to try the same idea here in our construction. That is, if $\mathbf{X}$ is the sum of two independent sources, then one can take two linear seeded extractors $\text{Ext}_1, \text{Ext}_2$ such that Ext_1 has large seed length, Ext_2 has seed length $O(\log n)$ and output length the same as the seed length of Ext_1, and compute $\text{Ext}_1(\mathbf{X}, \text{Ext}_2(\mathbf{X}, r))$ for every possible choice r of Ext_2's seed. However, the problem now is that the sampling procedure becomes correlated, and even with linear seeded extractors we do not know how to analyze it.

We thus turn to another approach, used by Li in his multi-source extractor [Li13a]. The idea is that, assume that $\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_C$ is the sum of some constant C number of independent sources (instead of just two independent sources). Then if we apply a linear seeded extractor to $\mathbf{X}$, by the property of the extractor for every fixed seed the output will also be the XOR of C independent outputs from each $\mathbf{X}_i$. If every output is ϵ-close to uniform for some error ϵ, then the error after the XOR will be reduced to roughly ϵ^C. Thus, if we take C to be a large enough constant, this error will be much smaller than $1/N$ where N is the number of rows in the somewhere

random source. At this point we can use a union-bound type argument to show that the somewhere random source is actually $N\epsilon^C = 1/\text{poly}(n)$-close to another somewhere random source where a large fraction of the rows are *truly uniform*. Thus we can switch to the new somewhere random source and only introduce an error of $1/\text{poly}(n)$.

8.3 The Extractor Construction

In this section we construct explicit extractors for (n, k, C)-sumset sources where $k = \text{polylog}(n)$ and C is a large enough constant.

Theorem 8.3.1 (Theorem 19 restated)**.** *There exists constants $c, C > 0$ and a small constant $\beta_1 > 0$ such that for all $n \in \mathbb{N}$, there exists a polynomial time computable extractor for $(n, k, C+1)$-sumset sources, $k \geq \log^c(n)$, with error $n^{-\Omega(1)}$ and output length k^{β_1}.*

We use the rest of the section to prove Theorem 8.3.1. We claim that the function computed by Algorithm 9 is the required extractor. We first set up the parameters and ingredients used by Algorithm 9.

- Let $\beta = 1/20$, $t = k^\beta, \epsilon = 1/n^2$.
- Let $c = (\lambda + 1)/\beta$.
- Let $\text{LExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^{n_1}$, $n_1 = \sqrt{k}$, be the linear seeded extractor from Theorem 2.1.6 set extract from min-entropy k with error ϵ. Thus $d = c_1 \log n$, for some constant c_1. Let $D = 2^d = n^{c_1}$.
- Let $C = c_1 + 2, k' = d^2, \epsilon_1 = 1/D^{2t} = 1/n^{2tc_1}, n_2 = k^{4\beta}, k'' = n_2^2 = k^{8\beta}, \delta = (2c_1 - 1)/2c_1$.
- Let $\text{LExt}_1 : \{0,1\}^{n_1} \times \{0,1\}^{d_1} \to \{0,1\}^{d_1}$ and $\text{LExt}_2 : \{0,1\}^{n_2} \times \{0,1\}^{d_1} \to \{0,1\}^{d_1}$ be instantiations of the linear seeded extractor from Theorem 2.1.5, both set to extract from min-entropy k' with error ϵ_1. Thus, $d_1 = O(\log^2(k/\epsilon_1)) = O(t^2 \log^2 n)$ and $d_2 = O(\log^2(k/\epsilon_1)) = O(t^2 \log^2 n)$. Finally let $\text{LExt}_3 : \{0,1\}^{n_1} \times \{0,1\}^{d_3} \to \{0,1\}^{n_2}$ be an instantiation of the linear

seeded extractor from Theorem 2.1.5 set to extract from min-entropy k'' with error ϵ_1. Thus, $d_3 = O(\log^2(n_1/\epsilon_1)) = O(t^2 \log^2 n)$. Let ACB be the function computed by Algorithm 4 using these linear seeded extractors.

- Let bitExt : $\{0,1\}^D \to \{0,1\}^m$, $m = t^\alpha$, be the extractor from Theorem 13 set to extract from (q, t, γ)-non-oblivious sources where $q = D^\delta$ and $\gamma = 1/D^{t+1}$.

Algorithm 9: SUMExt(x)
Input: A bit string $x = x_1 + \ldots + x_{C+1}$, where each x_i is a bit string of length n.
Output: A bit string of length m.

1 Let w be the $n_1 \times D$ boolean matrix whose i^{th} row w_i is given by LExt(x, s_i).
2 Let v be the $n_2 \times D$ boolean matrix whose i^{th} row v_i is given by ACB(w_i, x, s_i).
3 Let r be the first column of the matrix v. Output bitExt(r).

We prove the following claims about the random variables computed in Algorithm 9 from which Theorem 8.3.1 is direct.

Claim 8.3.2. $\mathbf{V}$ *is* $1/n^{O(1)}$*-close to a somewhere-random source* $\mathbf{V}'$ *containing a subset* R *of rows,* $|R| \geq D - D^\delta$ *such that the joint distribution of any* t *distinct rows in* R *is* γ*-close to* $\mathbf{U}_{tm}$*.*

Proof. Since LExt is a strong seeded extractor, it follows that for any $j \in [C]$, there exists a subset $S_j \subset \{0,1\}^d$, $|S_j| \geq (1 - \sqrt{\epsilon})D$, such that for any $s \in S_j$ LExt$(\mathbf{X}, s_j)$ is $\sqrt{\epsilon}$-close to $\mathbf{U}_{n_1}$. Thus, by a union bound, it follows that there exists a set $S \subset \{0,1\}^d$,

$$|S| \geq (1 - C\sqrt{\epsilon})D > D - D^\delta,$$

(the inequality follows by our choice of parameters) such that for any $s_i \in S$, LExt$(\mathbf{X}_j, s_i)$ is $\sqrt{\epsilon}$-close to $\mathbf{U}_{n_1}$ for each $j \in [C]$.

Since LExt is linear seeded, it follows that for any $i \in [D]$, it follows that $\mathbf{W}^i = \text{LExt}(\mathbf{X}, s_i) = \left(\sum_{j=1}^{C} \text{LExt}(\mathbf{X}_j, s_i)\right) + \text{LExt}(\mathbf{X}_{C+1}, s_i)$. Thus if $s_i \in S$, then by Lemma 2.3.8, $\left(\sum_{j=1}^{C} \text{LExt}(\mathbf{X}_j, s_i)\right)$ is $\epsilon^{C/2}$-close to $\mathbf{U}_{n_1}$. Using a hybrid argument, it follows that $\mathbf{W}$ is $D\epsilon^{C/2}$-close to a $D \times n_1$ matrix

$\overline{\mathbf{W}}$, whose i^{th} row $\overline{\mathbf{W}}^i$ is equal to $\mathbf{W}^i$ if $s_i \notin S$, and otherwise is given by $\mathbf{Y}^i + \mathrm{LExt}(\mathbf{X}_{C+1}, s_i)$, where $\mathbf{Y}^i$ follows the distribution $\mathbf{U}_{n_2}$. We note that the $\mathbf{Y}^i$'s can be arbitrarily correlated.

Thus, $\mathbf{V}$ is $D\epsilon^{C/2}$-close to a $D \times n_2$-matrix $\overline{\mathbf{V}}$ such that if $s_i \in S$, then the i^{th} row $\overline{\mathbf{V}}^i$ is given by $\mathrm{ACB}(\mathbf{Y}^i + \mathrm{LExt}(\mathbf{X}_{C+1}, s_i), \mathbf{X}, s_i)$.

Now consider any subset $\{s_{i_1}, \ldots, s_{i_t}\} \subset S$ of size t. We claim that

$$(\overline{\mathbf{V}}^{i_1}, \ldots, \overline{\mathbf{V}}^{i_t}) \approx_{O(td\epsilon)} \mathbf{U}_{tm}.$$

We fix the random variable $\{\mathrm{LExt}(\mathbf{X}_{C+1}, s_{i_1}), \ldots, \mathrm{LExt}(\mathbf{X}_{C+1}, s_{i_t})\}$. As a result of this fixing, $\mathbf{X}_{C+1}$ has min-entropy at least $k - tn_1 - \log(1/\epsilon) > k/2$ with probability at least $1 - \epsilon$. Let $\mathbf{Z} = \sum_{j=1}^{C} \mathbf{X}_j$.

Thus,

$$(\overline{\mathbf{V}}^{i_1}, \ldots, \overline{\mathbf{V}}^{i_t}) = (\mathrm{ACB}(\mathbf{Y}^1 + a_1, \mathbf{X}_{C+1} + \mathbf{Z}, s_{i_1}), \ldots, \mathrm{ACB}(\mathbf{Y}^t + a_t, \mathbf{X}_{C+1} + \mathbf{Z}, s_{i_t})),$$

where $a_1, \ldots, a_t$ are some constants.

We now invoke Theorem 3.4.2 noting that the following conditions hold by our choice of parameters:

- $\mathbf{X}_{C+1}$ is independent of $\{\mathbf{Z}, \mathbf{Y}^1, \ldots, \mathbf{Y}^t\}$.
- Each s_{i_g} is a distinct bit string of length d.
- $k/2 \geq k' + 8td_1d + \log(1/\epsilon)$.
- $n_2 \geq k' + 3td_1 + \log(1/\epsilon)$.
- $n_1 \geq k' + 10td_1d + (4td + 1)n_2 + \log(1/\epsilon)$.

Thus,

$$(\mathrm{ACB}(\mathbf{Y}^1 + a_1, \mathbf{X}_{C+1} + \mathbf{Z}, s_{i_1}), \ldots, \mathrm{ACB}(\mathbf{Y}^t + a_t, \mathbf{X}_{C+1} + \mathbf{Z}, s_{i_t})) \approx_{O(dt\epsilon_1)} \mathbf{U}_{tm}.$$

We note that by our choice of parameters, the following inequalities hold:

- $dt\epsilon_1 < 1/D^{t+2}$.
- $\epsilon^{C/2}D \leq 1/n^2$.

The claim now follows from the fact the above argument holds for any arbitrary size t subset of S and the fact that $\mathbf{V}$ is $\epsilon^{C/2}D$-close to $\overline{\mathbf{V}}$. □

Claim 8.3.3. *$\mathbf{V}'$ is $1/n^{O(1)}$-close to $\mathbf{U}_m$.*

Proof. Follows directly from Claim 8.3.2 and Theorem 13. □

Remark 8.3.4. *It is not hard to see that the results in this section generalize to sumset sources over any field, i.e., sources of the form $\mathbf{X} = \sum_{i=1}^{C} \mathbf{X}_i$, where each $\mathbf{X}_i$ is a source on $\mathbb{F}_q^n$ for some prime power q, where for $a, b \in \mathbb{F}_q^n$, $a + b$ denotes the standard vector addition.*

Chapter 9

Extractors for Small-Space Sources

[1] Trevisan and Vadhan [TV00] introduced the problem of constructing seedless extractors for the class of samplable sources, where the weak random source is generated by a computationally bounded algorithm. They constructed explicit extractors for such sources based on strong but plausible complexity-theoretic assumptions. Subsequently, Kamp et al. [KRVZ11] studied the problem of constructing seedless extractors for small-space sources, where the weak source is generated by a small width branching program. We define this model more formally below.

Definition 9.0.1. *[KRVZ11] A space s source $\mathbf{X}$ is generated by taking a random walk on a branching program of length n and width 2^s, where each edge of the branching program is labelled with a transition probability and a bit. Thus a bit of the source is generated for each step taken on the branching program, and the source X is the concatenation of all the bits.*

As observed in [KRVZ11], the model of small space sources generalizes many previously studied sources, including von Neumann's source of independent coin flips with unknown bias [vN51], the finite Markov chain model studied by Blum [Blu86], a generalization of bit-fixing sources known as symbol-fixing sources [KZ07b], and sources consisting of many independent sources. However, the class of affine sources appears not to be related to small space sources.

[1] parts of this chapter have been previously published [CL16b]

Using the probabilistic method, one can show that error ε extractors exist for space s sources with min-entropy $k \geq 2s + \log s + O(\log(n/\epsilon))$. However, previously the best known explicit extractor for space s sources is from the work of Kamp et al. [KRVZ11], which requires min-entropy $k \geq \gamma n$ and space $s \leq \gamma^3 n$, where $\gamma > n^{-\delta}$ for some small universal constant δ. In other words, their extractor requires almost linear min-entropy even for sources with space as small as 1, while we know from the probabilistic method that for space $O(\log n)$ sources one can hope to construct extractors for min-entropy $O(\log n)$. In addition, the techniques used in [KRVZ11] start out by reducing to the so called *total-entropy independent sources*, and it can be shown that this reduction has a fundamental bottleneck and cannot possibly go below min-entropy $\sqrt{n}$.

9.1 Our Result and Overview of Techniques

The results in this chapter are based on joint work with Xin Li [CL16b].

We show how to extract from space s sources when $k \geq 2^{\log^{0.5+\alpha}(n)} s^{1+10\alpha}$, for any constant $\alpha > 0$. Thus for $s = n^{o(1)}$, we only need min-entropy $n^{o(1)}$. This significantly improves previous results in terms of min-entropy requirement, and in particular break the $\sqrt{n}$ min-entropy barrier.

Theorem 20. *For any constant $\alpha > 0$ and for all $n, k, s \in \mathbb{N}$ with $k \geq 2^{\log^{0.5+\alpha}(n)} s^{1+10\alpha}$, there exists a polynomial time computable extractor for space s sources on n bits with min-entropy at least k, with error $n^{-\Omega(1)}$ and output length k^{α}.*

We obtain our result by showing a reduction from the task of extracting from small-space sources to the problem of extracting from sumset sources. We briefly describe the reduction below and refer the reader to Section 9.2 for more details. Our extractor follows immediately from the reduction.

Note that as observed in [KRVZ11], if we partition a small space source into several blocks, and condition on the event that the branching program generating the source reaches some specific vertices at the end of each block, then the small space source becomes a convex combination of independent sources. This conditioning reduces the min-entropy of the source, but since the

branching program has small width we would expect that there is still much entropy left. However, the problem is that the entropy could now be distributed in these blocks in some arbitrary way, with the only guarantee being a lower bound on the total amount of entropy. This is referred to as a *total entropy source* as in [KRVZ11]. The problem with the approach in [KRVZ11] is that one has to use a fixed partition of the source, so that the blocks can be used as inputs to an extractor for independent sources. This introduced a bottleneck of entropy $\sqrt{n}$, since if the block size is smaller than $\sqrt{n}$ then it could be the case that each block has entropy 1, while if the block size is larger than $\sqrt{n}$ then it could be the case that all entropy is concentrated in just one block.

We get around this obstacle by not relying on a fixed partition of the source. Instead, we show that when the min-entropy satisfies $k \geq 2^{\log^{0.5+\alpha}(n)} s^{1+10\alpha}$, the small space source is actually $2^{-k^{\Omega(1)}}$-close to a convex combination of (n, k^α, C)-sumset sources. On a high level, we show this reduction as follows. We first partition the small space source into some $\ell \gg C$ blocks with $\ell s \ll k$, and we condition on the fixing of the states of the random walk at the end of each block. This leaves us ℓ independent blocks such that their total min-entropy is roughly $k - \ell s$. Now if for some particular fixing, there are at least C blocks with min-entropy at least k^α, then under this fixing the source is an (n, k^α, C)-sumset source. If not, then our key observation is that most of the entropy (indeed, $k - \ell s - \ell k^\alpha = k - o(k)$ entropy) will be concentrated in at most $C - 1$ blocks. Therefore at least one block has min-entropy $(k - o(k))/(C - 1)$. Thus, for this block the entropy rate will be increased by a factor of roughly ℓ/C. We can then fix all other blocks and repeat the argument for this block. Specifically, we further divide the block into ℓ blocks and condition on the fixing of the intermediate states. Then for any particular fixing, either it is an (n, k^α, C)-sumset source or the entropy rate of one block gets increased again by a factor of ℓ/C. We note that the entropy rate cannot be larger than 1, so we know at some point it has to be an (n, k^α, C)-sumset source. Therefore the original source is a convex combination of sumset sources. Notice here the partitions are not fixed, but rather can be different for different fixings of the states.

We also consider a generalization of small space sources, where the underlying branching program produces bits of the source in an unkown (but oblivious) order. This is discussed in Section

9.3. We also obtain new results on extracting from total-entropy sources and somewhere entropy sources (see Section 9.4).

9.2 A Reduction from Small-Space Sources to Sumset Sources

In this section, we show that a small-space source is close to a convex combination of sumset sources. The idea is argue that either partitioning the source leads to a sumset source or results in increase in min-entropy rate of one of the partitions. Thus by repeating this argument, it must be that at some point we reach a sumset source, since otherwise we end up with a source with min-entropy rate more than 1.

Lemma 9.2.1. *For any constant $\alpha > 0$ and any constant integer $C \geq 2$, any space s source on n bits with min-entropy $k \geq 2^{\log^{0.5+\alpha}(n)} s^{1+10\alpha}$ is $2^{-k^{\Omega(1)}}$-close to a convex combination of (n, k', C)-sumset sources, where $k' = k^{\alpha}$.*

We note that Theorem 20 now directly follows from the explicit sumset extractor constructed in Chapter 8 (Theorem 19) and Lemma 9.2.1.

Proof of Lemma 9.2.1. Let $\ell = k^{\alpha/2}, \epsilon_1 = 2^{-k^{\alpha}}, k_{th} = k^{\alpha}$ be fixed parameters that we set with foresight. Let $\mathbf{X}$ be a space s source on n bits with min-entropy at least k. We partition $\mathbf{X}$ into ℓ equi-sized blocks of length $n_1 = n/\ell$. Let $\mathbf{X}_i$, denote the i'th block where $i \in [\ell]$ (thus $\mathbf{X}_i$ is a source on n/ℓ bits). We now condition on the initial state of small-space branching program at each of these blocks, and let k_i denote the min-entropy in $\mathbf{X}_i$ after this conditioning. Observe that $\mathbf{X}_i$'s are now independent sources. It follows from Lemma 2.3.7 that with probability at least $1 - \epsilon_1$,

$$\sum_{i=1}^{\ell} k_i \geq k - \ell s - \log(1/\epsilon_1). \tag{9.1}$$

Consider any such good fixing of the states such that the above inequality holds. The proof now goes via analysing two cases. Since we iterate this argument, each time with a new source, let

$\mathbf{X}^1 = \mathbf{X}$ and $k_{(1)} = k$.

Case 1: $|\{i \in [\ell] : k_i \geq k_{th}\}| \geq C$. The proof is direct in this case. For simplicity, suppose $\mathbf{X}_1, \ldots, \mathbf{X}_C$ each have min-entropy at least k^α. We fix the sources $\mathbf{X}_{C+1}, \ldots, \mathbf{X}_\ell$. Now, for each $i \in [C]$, define the source $\mathbf{Y}_i$ on n bits whose projection onto the the i'th block is $\mathbf{X}_i$ and the rest of the co-ordinates are fixed to 0. It follows that $\mathbf{X} = \eta + \sum_{i=1}^{C} \mathbf{Y}_i$ (for some constant string $\eta \in \{0,1\}^n$) and hence is a (n, k', C)-sumset source. Thus $\mathbf{X}$ is at distance at most ϵ_1 from a convex combination of such sumset sources.

Case 2: $|\{i \in [\ell] : k_i \geq k_{th}\}| < C$. Using (9.1), it follows that there exists distinct $C-1$ partitions, say $i_1, \ldots, i_{C-1}$ such that

$$\sum_{j=1}^{C-1} k_{i_j} \geq k_{(1)} - \ell(s + k^\alpha) - \log(1/\epsilon_1).$$

Thus, by an averaging argument, it follows that there exists some $j \in [C-1]$, such that

$$k_{i_j} \geq \frac{k_{(1)} - \ell(s + k^\alpha) - \log(1/\epsilon_1)}{C-1}.$$

Hence the source $\mathbf{X}_{i_j}$ (on $n_1 = n/\ell$ bits) has min-entropy rate

$$\frac{k_{(1)} - \ell(s + k^\alpha) - \log(1/\epsilon_1)}{C-1} \cdot \frac{\ell}{n}$$

Thus, using the fact that $k_{(1)} > (sk^{\alpha/2} + 2k^{\alpha+\alpha})^{1+\alpha}$, the min-entropy rate of $\mathbf{X}_{i_j}$ is at least $\frac{k_{(1)}\ell}{2nC}$, and hence

$$\frac{H_\infty(\mathbf{X}_{i_j})}{n_1} \geq \frac{\ell}{2C} \cdot \frac{H_\infty(\mathbf{X}^1)}{n}.$$

We now repeat the argument (i.e, analyzing the Cases 1 and 2) with $\mathbf{X}^1$ replaced by $\mathbf{X}^2 = \mathbf{X}_{i_j}$ (and we fix all other sources). However, for different iterations of the argument, we do not change values of the parameters ℓ, ϵ, k_{th}, and they are fixed to k^α, 2^{-k^α} and k^α respectively, where $k = H_\infty(\mathbf{X})$.

Suppose, if possible, that for h iterations of this argument, each time we end up in Case 2.

Thus, we now have a source $\mathbf{X}^h$ on n/ℓ^h bits with min-entropy rate at least $(\frac{\ell}{2C})^h \cdot \frac{k}{n}$. To derive a contradiction using the fact that the min-entropy rate is at most 1, we require

- $\left(\frac{\ell}{2C}\right)^h \cdot \frac{k}{n} > 1$,
- $\frac{n}{\ell^h} \geq k^\alpha$
- $\frac{k}{(2C)^h} > (sk^{\alpha/2} + 2k^{2\alpha})^{1+\alpha}$.

(The first condition is to ensure that the min-entropy rate is more than 1, the second condition ensures that the length of the source $\mathbf{X}^h$ is large enough and finally the third condition is a lower bound the min-entropy of $\mathbf{X}^h$, which is required when we apply our argument on $\mathbf{X}^{h-1}$.)

Pick $h = 1 + \frac{\log n - \log k}{\log \ell - \log(2C)}$. It is easy to check that the first condition holds. Further the second and third conditions follow from the fact that $k > s^{1+10\alpha} 2^{\log^{0.5+\alpha}(n)}$. Thus, it must be that in at most h iterations of the argument, we are in Case 1 and hence $\mathbf{X}$ is close to a convex combination of (n, k', C)-sumset sources. We note that the statistical distance to the convex combination is at most $O\left(\epsilon_1 \frac{\log n}{\log k}\right)$. □

9.3 Any-Order Small-Space-Sources

Consider the following natural generalization of small-space sources.

Definition 9.3.1 (Any-Order-Small-Space-Sources)**.** *An any-order-space s source X on $[r]^n$ is generated by an r-way branching program of length n and width 2^s and a permutation $t : [n] \to [n]$ in the following way: The r-way branching program is a layered graph with $n+1$ layers and a single start vertex. Each edge is labeled with a variable X_j, a probability value and a symbol in $[r]$. Further all edges between the ith and $(i+1)$th layer are labelled with same variable $X_{t(i)}$. The output of the source is a random walk starting from the start vertex, assigning the symbol on the edge to the corresponding variable and finally outputting the generated string.*

It is easy to see that the reduction in the above section generalizes to the class of any-order small-space sources. Thus, we have the following theorem.

Theorem 21. *For any constant $\alpha > 0$ and for all $n, k, s \in \mathbb{N}$ with $k \geq 2^{\log^{0.5+\alpha}(n)} s^{1+10\alpha}$, there exists a polynomial time computable extractor for the class of any-order space s sources on n bits with min-entropy at least k, with error $n^{-\Omega(1)}$ and output length k^{α}.*

9.4 Total Entropy and Some-Where Entropy Sources

As an intermediate model to extract from small space sources, [KRVZ11] introduced the above mentioned *total entropy independent sources*. This is a collection of r independent sources of length ℓ such that the total min-entropy of all r sources is at least k. By the probabilistic method, one can show that error ϵ extractors exist for total min-entropy k independent sources as long as $k \geq \max\{\ell, \log\log(r/\epsilon)\} + \log r + 2\log(1/\epsilon) + O(1)$.[2] Essentially, k can be as small as $\ell + \log r + O(1)$. However, the best known extractors in [KRVZ11] are far from this. Specifically, the extractors there need to have either $k \geq \Omega(r\ell)$ or $k \geq (2^{\ell}\log r)^{C}$ for some constant $C > 1$.

We substantially improve these results by constructing a new extractor that only requires min-entropy $O(\ell) + \text{polylog}(r\ell)$, which comes close to the probabilistic bound. In particular, we have the following result.

Theorem 9.4.1. *There exist constants $c, C > 0$ and a small constant $\beta > 0$ such that for all $r, \ell, k \in \mathbb{N}$ with $k \geq C(\ell + \log^{c}(r\ell))$, there exists a polynomial time computable extractor for r independent sources over $\{0,1\}^{\ell}$ with total min-entropy k, with error $(r\ell)^{-\Omega(1)}$ and output length k^{β}.*

To prove the theorem we show the following lemma.

Lemma 9.4.2. *For any $t, C \in \mathbb{N}$, let $\mathbf{X}_1, \cdots, \mathbf{X}_r \in (\{0,1\}^{\ell})^{r}$ be r independent sources over $\{0,1\}^{\ell}$*

[2]Note that $k > \ell$ is necessary, otherwise the entropy could be contained in just one source, making extraction impossible.

with total min-entropy $k \geq C(\ell + t)$. Then there exists a partition of the r sources into C disjoint subsets $\mathbf{Y}_1, \cdots, \mathbf{Y}_C$ such that each Y_i has min-entropy at least t.

Proof. We prove the lemma by induction on C. For the case where $C = 1$, one can view the whole set $\mathbf{X}_1, \cdots, \mathbf{X}_r$ as a partition Y_1, and it is clear that Y_1 has min-entropy $k \geq \ell + t > t$. Now suppose the lemma holds for some $C \in \mathbb{N}$, we show that it holds for $C + 1$.

First notice that for two independent sources $\mathbf{X}, \mathbf{Y}$, we have that $H_\infty(\mathbf{X} \circ \mathbf{Y}) = H_\infty(\mathbf{X}) + H_\infty(\mathbf{Y})$. Now, consider the smallest i such that $\mathbf{X}_1 \circ \cdots \circ \mathbf{X}_i$ has min-entropy at least t. We know such an i exists because $\mathbf{X}_1 \circ \cdots \circ \mathbf{X}_r$ has min-entropy at least $k \geq (C+1)(\ell + t) > t$. Since i is the smallest, we know that $\mathbf{X}_1 \circ \cdots \circ \mathbf{X}_{i-1}$ has min-entropy at most t. Note that $\mathbf{X}_i$ has min-entropy at most ℓ, thus $\mathbf{X}_1 \circ \cdots \circ \mathbf{X}_i$ has min-entropy at most $t + \ell$. Next, since $\mathbf{X}_1 \circ \cdots \circ \mathbf{X}_r$ has min-entropy at least $k \geq (C+1)(\ell + t)$, we know that $\mathbf{X}_{i+1} \circ \cdots \circ \mathbf{X}_r$ has min-entropy at least $k - (t + \ell) = C(t + \ell)$. Now we can apply the induction hypothesis and we see that there exists a partition of $\mathbf{X}_{i+1} \cdots \mathbf{X}_r$ into C disjoint subsets such that each subset has min-entropy at least t. Put in $\mathbf{X}_1 \circ \cdots \circ \mathbf{X}_i$ we get $C + 1$ disjoint subsets. □

By setting $t = \log^c(r\ell)$ and combining the lemma with Theorem 19, we immediately obtain Theorem 9.4.1.

In order to extract from total entropy independent sources, [KRVZ11] actually argues that since the total entropy is at least k, some of the independent sources will have entropy at least k' (the relation between k and k' depends on the number of sources). Therefore, total entropy sources reduce to independent sources where some of them have a certain amount of min-entropy. We call such sources *somewhere entropy independent sources.*

Definition 9.4.3. *An (n, k, ℓ)-somewhere-C source consists of ℓ independent sources $\mathbf{X}_1, \ldots, \mathbf{X}_\ell$, each on n bits, such that at least C of the $\mathbf{X}_i$'s have min-entropy at least k.*

Note that C here needs to be at least 2. In this context, our extractor for sumset sources from Theorem 19 actually gives an extractor for an (n, k, ℓ)-somewhere-C source with $k \geq \log^c n$

for some constants $C, c > 1$, and outputs $k^{\Omega(1)}$ bits. Note that the number of sources ℓ here is irrelevant since we can just take the sum of the sources and fix any other source that does not have min-entropy k.

In fact, we can use a simpler method to get a slightly stronger result. We show that we can extract from (n, k, ℓ)-somewhere 2 sources for $k = \text{polylog}(n)$ and any integer ℓ (with the extractor running in time $\text{poly}(n, \ell)$).

Theorem 9.4.4. *There exists a constant $c > 0$ and a small constant $\beta > 0$ such that for all $n, k, \ell \in \mathbb{N}$ with $k \geq \log^c n$, there exists an extractor computable in time $\text{poly}(n, \ell)$ for (n, k, ℓ)-somewhere-2 sources, with error $n^{-\Omega(1)}$ and output length $\Omega(k)$.*

Proof. Let $2\text{Ext} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$, $m = k/10$ be the 2-source extractor from Theorem 14 set to extract from min-entropy $k/2$ with error $\epsilon = 1/n^{\Omega(1)}$. Define the function $\text{Ext} : \{0,1\}^{\ell n} \to \{0,1\}^m$ as

$$\text{Ext}(x_1, \ldots, x_\ell) = \sum_{1 \leq i < j \leq \ell} 2\text{Ext}(x_i, x_j).$$

We claim that for any (n, k, ℓ)-somewhere 2-source $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_\ell\}$,

$$|\text{Ext}(\mathbf{X}) - \mathbf{U}_m| \leq \epsilon.$$

We prove this in the following way. Since the function Ext is symmetric, we can assume without loss of generality that the sources $\mathbf{X}_1$ and $\mathbf{X}_2$ have min-entropy at least k each. Fix the sources $\mathbf{X}_3, \ldots, \mathbf{X}_\ell$. Thus, after this fixing

$$\text{Ext}(\mathbf{X}_1, \mathbf{X}_2, x_3 \ldots, x_\ell) = 2\text{Ext}(\mathbf{X}_1, \mathbf{X}_2) + \left(\sum_{j=3}^{\ell} 2\text{Ext}(\mathbf{X}_1, x_j)\right) + \left(\sum_{j=3}^{\ell} 2\text{Ext}(\mathbf{X}_2, x_j)\right) + s,$$

for some constant $s \in \{0,1\}^m$. Now, we observe that $\mathbf{A} = \left(\sum_{j=3}^{\ell} 2\text{Ext}(\mathbf{X}_1, x_j)\right)$ is a random variable on $\{0,1\}^m$ and is deterministic function of $\mathbf{X}_1$. Thus, we fix $\mathbf{A}$, and using Lemma 2.3.7, $\mathbf{X}_1$ has min-entropy at least $0.9k - m$ with probability $1 - 2^{-k^{0.1}}$. Similarly, $\mathbf{B} = \left(\sum_{j=3}^{\ell} 2\text{Ext}(\mathbf{X}_2, x_j)\right)$

is a random variable on $\{0,1\}^m$ and is deterministic function of $\mathbf{X}_2$. Thus, we fix $\mathbf{B}$, and $\mathbf{X}_2$ has min-entropy at least $0.9k - m$ with probability $1 - 2^{-k^{0.1}}$. Thus, after this fixing

$$\mathrm{Ext}(\mathbf{X}) = 2\mathrm{Ext}(\mathbf{X}_1, \mathbf{X}_2) + s',$$

for some constant $s' \in \{0,1\}^m$. Further $\mathbf{X}_1$ and $\mathbf{X}_2$ are still independent, each with min-entropy at least $0.8k$ (with probability at least $1 - 2^{-k^{\Omega(1)}}$). The result now follows since 2Ext is a 2-source extractor for min-entropy $k/2$. □

Chapter 10

Extractors for Interleaved Sources

[1]

Raz and Yehudayoff [RY11] introduced a natural generalization of the class of independent sources, which we call *interleaved sources*. To formally define this class of sources, we introduce some notation. Let $\sigma : [n] \to [n]$ be any permutation. For any string $w \in \{0,1\}^n$, define the string $s = w_\sigma \in \{0,1\}^n$ such that $s_{\sigma(i)} = w_i$ for $i = 1, \ldots, n$.

Definition 10.0.1 (Interleaved Sources)**.** *Let $\mathbf{X}_1, \ldots, \mathbf{X}_C$ be independent (n,k)-sources on $\{0,1\}^n$ and let $\sigma : [Cn] \to [Cn]$ be any permutation. Then $\mathbf{Z} = (\mathbf{X}_1 \circ \ldots \circ \mathbf{X}_C)_\sigma$ is an (n,k,C)-interleaved source.*

Besides being a natural generalization of independent sources, the original motivation for studying these sources came from an application found by Raz and Yehudayoff [RY11] in proving lower bounds for arithmetic circuits. Further, such extractors give examples of explicit functions with high best-partition communication complexity.

Using the probabilistic method, one can show that extractors exist for (n,k,C)-interleaved sources with $C = 2$ and $k = O(\log n)$. However the known constructions are far from this in terms of entropy requirement. The construction in [RY11] works for (n,k,C)-interleaved sources

[1] parts of this chapter have been previously published [CZ16b]

for $k > (1-\delta)n$ and $C = 2$, where δ is a small constant.

10.1 Our Results and Applications

The results in this chapter are based on joint works with Xin Li and David Zuckerman [CZ16b, CL16b].

Note that an (n, k, C)-interleaved source is also a special case of an (n, k, C)-sumset source, by naturally extending each source in the definition to have bits 0 in all other positions. Using our extractor for sumset sources from Chapter 8, we thus substantially improve previous results in terms of min-entropy requirement for a large enough constant C. In particular, we obtain explicit extractors that work for the interleaving of a constant number of independent sources, each with polylogarithmic min-entropy.

Theorem 22. *There exist constants $c, C > 0$ and a small constant $\beta > 0$ such that for all $n, k \in \mathbb{N}$ with $k \geq \log^c n$, there exists a polynomial time computable extractor for (n, k, C)-interleaved sources, with error $n^{-\Omega(1)}$ and output length k^β.*

Proof. Suppose $\mathbf{X}$ on Cn is an interleaving of the independent sources $\mathbf{X}_1, \ldots, \mathbf{X}_C$ (each on n bits). Define independent sources $\mathbf{Y}_1, \ldots, \mathbf{Y}_C$, each on Cn bits, such that $\mathbf{Y}_i$ matches $\mathbf{X}$ on the co-ordinates belonging to the source $\mathbf{X}_i$, and $\mathbf{Y}_i$ is fixed to 0 everywhere else. Hence $\mathbf{X} = \sum_1^C \mathbf{Y}_i$ and thus, $\mathbf{X}$ is a (Cn, k, C)-sumset source. The result now follows from Theorem 19. □

However note that this does not yield extractors for $(n, k, 2)$-interleaved sources since C (in the above theorem) is a large constant.

To extract from $(n, k, 2)$ sources, we develop a simple technique that yields explicit extractors that work for lower min-entropy rates. In particular, our method yields explicit extractors for min-entropy rate 0.51 for two interleaved sources, when the sources are over a finite field of large enough (constant) characteristic.

We show how to convert any two-source extractor that is a function of the sum of its inputs into an extractor for a 2-interleaved source. Our method of converting a two-source extractor into an extractor for interleaved sources is based on explicit constructions of certain combinatorial sets, which we call (r, s)-spanning sets. These spanning sets are essentially subspace-evasive sets with different parameters than studied earlier (see Section 10.2.1 for more details). It turns out that the columns of parity check matrices of linear codes with good erasure list-decodability form spanning sets with good parameters. We discuss this in detail later.

Next, we observe that an existing two-source extractor from [CG88] is a function of the sum of the inputs. This leads to our construction of an extractor for 2-interleaved sources with one source having min-entropy at least $(1-\alpha)n$ and the other source having min-entropy at least $\lambda \log n$ (for some $\alpha, \lambda > 0$). In particular, we have the following theorem.

Theorem 23. *For some $\delta > 0$ and any $\lambda > 0$, there exists an explicit function $\mathrm{ext} : \{0,1\}^{2n} \to \{0,1\}^m$, $m = \lambda \log n$, such that if $\mathbf{X}$, $\mathbf{Y}$ are independent sources on $\mathbb{F}_2^n$ with min-entropy k_1, k_2 respectively satisfying $k_1 > (1-\delta)n$ and $k_2 > 35 \max\{\log n, m\}$, $t : [2n] \to [2n]$ is any permutation, then*

$$|\mathrm{ext}((\mathbf{X} \circ \mathbf{Y})_t) \circ \mathbf{X} - U_m \circ \mathbf{X}| = n^{-\Omega(1)}.$$

Next, we show that for any large enough constant prime p, if the 2-interleaved source is on $[p]^{2n}$, we can extract when one source has min-entropy rate at least 0.51 and the other source has min-entropy rate at least $c \log n/n$.

Theorem 24. *There exists $c > 0$ such that for any $\delta, \lambda > 0$ and any prime $p > 2^{\frac{c}{\delta}}$, there exists an explicit function $\mathrm{ext}_p : \mathbb{F}_p^{2n} \to \{0,1\}^m$, $m = \lambda \log n$, such that if $\mathbf{X}$ and $\mathbf{Y}$ are independent sources on $\mathbb{F}_p^n$ with min-entropy k_1, k_2 respectively, satisfying $k_1 > (\frac{1}{2}+\delta)n \log p$ and $k_2 > \frac{5}{\delta} \max\{\log n \log p, m\}$, $t : [2n] \to [2n]$ is any injective map, then*

$$|\mathrm{ext}_p((\mathbf{X} \circ \mathbf{Y})_t) \circ \mathbf{X} - U_m \circ \mathbf{X}| = n^{-\Omega(1)}.$$

We give various related constructions achieving different tradeoffs between min-entropy, error, and output length. This is summarized in Table 10.1.

We show that random sets are (r, s)-spanners with high probability (see Lemma 10.3.10). By our proof technique, any improved construction of an (r, s)-spanning set matching the probabilistic method will yield extractors for 2-interleaved sources on $\{0,1\}^{2n}$ that have essentially the same min-entropy requirement as the standard (non-interleaved) setting.

10.1.1 Best-Partition Communication Complexity

Since Yao introduced communication complexity in 1978 Yao [Yao79], there has been an extensive amount of research done on various models of communication (see [KN97] for formal definitions and background). We recall the definition of the randomized best-partition communication complexity of an arbitrary function $f : [R]^{2n} \to \{0,1\}$, which generalizes the usual setting where the partition of inputs is known.

Let Alice and Bob be two players who want to collectively compute f following a protocol Π and having access to a common random string r. Fix an arbitrary partition of the set $[2n]$ into 2 subsets of equal size, say S and T. For arbitrary $x, y \in [R]^n$, Alice is given x and Bob receives y and the goal is to compute $f(z)$ with probability at least $1 - \epsilon$, where $z \in [R]^{2n}$ such that $z_S = x$ and $z_T = y$.

For any protocol Π, the randomized communication cost of f with respect to an equi-partition $S, T \subset [2n]$ denoted by $R^{\epsilon}_{\Pi,S,T}(f)$, is defined to be the maximum communication between Alice and Bob over all inputs x, y in the scenario described above. The best-partition communication complexity of f, denoted by $R^{best,\epsilon}(f)$ is defined as:

$$R^{best,\epsilon}(f) = \min_{\Pi}\{\min_{\substack{S,T:|S|=|T|=n,\\ S\cup T=[2n]}} R^{\epsilon}_{\Pi,S,T}(f)\}.$$

Lower bounds on the best-partition communication complexity of f implies lower bounds

on branching programs computing f [AM86] and also imply time/space tradeoffs for VLSI circuits [Len90].

Raz and Yehudayoff [RY11] proved the following lower bound.

Theorem 10.1.1 ([RY11])**.** *For some $\beta > 0$, there exists an explicit function $f : \{0,1\}^{2n} \to \{0,1\}$ such that the randomized best-partition communication complexity of f with error $\epsilon = \frac{1}{2} - 2^{-\beta n}$ is at least βn.*

The constant β in the above theorem is, however, extremely small and arises from arguments in additive combinatorics. A similar bound also follows from their work for inputs on $[R]^{2n}$ (for any constant R) and it appears nontrivial to use their techniques to obtain bounds for larger β.

Our Results

We obtain the following result.

Theorem 25. *There exists $c > 0$ such that for any $\delta, \gamma > 0$ and any prime $p > 2^{\frac{c}{\delta}}$, there exists an explicit function $f : [p]^{2n} \to \{0,1\}$ such that the randomized best-partition communication complexity of f with error $\epsilon = \frac{1}{2} - p^{-\gamma n}$ is at least $(\frac{1}{4} - \delta - \gamma) n \log p$.*

We prove this using a well known technique of lower bounding randomized communication complexity by discrepancy. Our explicit function is the 1-bit extractor constructed in Theorem 10.4.7. However, we need to analyze the error of the extractor more carefully to obtain the above bound. We prove Theorem 25 in Section 10.6.

10.1.2 Interleaved Non-Malleable Extractors

We introduce the natural generalization of non-malleable extractors in the interleaved model.

We first recall the definition of a non-malleable extractor.

Definition 10.1.2 (Non-Malleable Extractor)**.** *A function* nmExt $: [R]^{2n} \to \{0,1\}^m$ *is a non-malleable extractor for min-entropy k and error ϵ if the following holds: If $\mathbf{X}$ is a source (on $[R]^n$)*

with min-entropy k, and $f : [R]^n \to [R]^n$ is any function with no fixed points, then

$$|\mathrm{nmExt}(\mathbf{X} \circ U_{[R]^n}) \circ \mathrm{nmExt}(\mathbf{X} \circ f(U_{[R]^n})) \circ U_{[R]^n} - U_m \circ \mathrm{nmExt}(\mathbf{X} \circ f(U_{[R]^n}) \circ U_{[R]^n}| \leq \epsilon.$$

The first explicit construction of a non-malleable extractors was given in [DLWZ14], with subsequent improvements of parameters achieved in [CRS12, Li12b]. However these constructions require min-entropy $> 0.49n$. In a recent work [CGL16], the min-entropy required was improved to $O(\log^2 n)$.

We initiate the study of non-malleable extractors in the interleaved model, where the extractor is guaranteed to work even when symbols from the source $\mathbf{X}$ and tampered seed $U_{[R]^n}$ arrive to the non-malleable extractor in a fixed but unknown interleaved order.

We formally define interleaved non-malleable extractors.

Definition 10.1.3 (Interleaved Non-Malleable Extractor)**.** *A function* $\mathrm{nmExt} : [R]^{2n} \to \{0,1\}^m$ *is a non-malleable extractor in the any-order model for min-entropy k and error ϵ if the following holds: If $\mathbf{X}$ is a source (on $[R]^n$) with min-entropy k, $f : [R]^n \to [R]^n$ is any function with no fixed points and $t : [2n] \to [2n]$ is any permutation, then*

$$|\mathrm{nmExt}((\mathbf{X} \circ U_{[R]^n})_t) \circ \mathrm{nmExt}((\mathbf{X} \circ f(U_{[R]^n}))_t) \circ U_{[R]^n} - U_m \circ \mathrm{nmExt}((\mathbf{X} \circ f(U_{[R]^n}))_t) \circ U_{[R]^n}| \leq \epsilon,$$

where U_m is independent of $U_{[R]^n}$.

In the above definition, when the seed has some min-entropy instead of being uniform, we say that the interleaved non-malleable extractor is weak-seeded.

Our Results

We give the first explicit construction of an interleaved non-malleable extractor. Further our non-malleable extractor is weak-seeded.

Theorem 26. *There exists $\lambda > 0$ such that for any $\delta > 0$, $c > c(\delta)$ and any prime $p > 2^{\frac{\lambda}{\delta}}$, there exists an explicit function* $\mathrm{nmExt} : \mathbb{F}_p^{2n} \to \{0,1\}^m$, *$m = O(\log n)$, such that if $\mathbf{X}$, $\mathbf{Y}$ are independent sources on $\mathbb{F}_p^n$ with min-entropy k_1, k_2 respectively, satisfying $k_1 > (\frac{1}{2} + \delta)n \log p$ and $k_2 > cm$, $t : [2n] \to [2n]$ is any permutation and $f : \mathbb{F}_p^n \to \mathbb{F}_p^n$ is any function with no fixed points, then*

$$|\mathrm{nmExt}((\mathbf{X} \circ \mathbf{Y})_t) \circ \mathrm{nmExt}((\mathbf{X} \circ f(\mathbf{Y}))_t) \circ \mathbf{Y} - U_m \circ \mathrm{nmExt}((\mathbf{X} \circ f(\mathbf{Y}))_t) \circ \mathbf{Y}| = n^{-\Omega(1)}.$$

As before, if we are allowed to run the non-malleable extractor in sub-exponential time, we can extract $\Omega(n)$ bits with error $2^{-\Omega(n)}$. See Theorem 10.5.4 for more details.

10.2 Outline of Constructions

10.2.1 Extractors for 2-Interleaved Sources

Our extractor for interleaved sources exploits the existence of good 2-source extractors which are functions of $\mathbf{X} + \mathbf{Y}$. To do this, we encode our source in a new way. Our encoding is based on explicit constructions of certain combinatorial sets, which we call spanning vectors.

Definition 10.2.1. *A set of vectors $S \subseteq \mathbb{F}_p^{\bar{\ell}}$ is (r, s)-spanning if the span of any r vectors of S has dimension at least s.*

Note that this is the same as a subspace-evasive set: Any $(s-1)$-dimensional subspace contains at most $(r-1)$ vectors in the set. However our parameters are quite different than studied previously [Gur11, DL12].

Our explicit constructions of spanning vectors are based on using the columns of a parity check matrix of a linear codes with good erasure list-decodability. Informally, an (e, L)-erasure list-decodable code $\mathcal{C}$ satisfies the property that at most L codewords agree on any particular subset of coordinates of size $n - e$. This property can then be used to lower bound the rank of any subset of e columns of the parity check matrix of $\mathcal{C}$. We refer the reader to Section 10.3 for more details.

We define the following encoding based on spanning vectors.

Definition 10.2.2. *For any (r,s)-spanning set $S = \{v_1, \ldots, v_\ell\} \subseteq \mathbb{F}_p^{\bar{\ell}}$ of size ℓ, the function* $\mathrm{enc} : \mathbb{F}_p^{\ell} \to \mathbb{F}_p^{\bar{\ell}}$ *defined as*

$$\mathrm{enc}(z) = \sum_{i=1}^{\ell} z_i v_i$$

is called an (r,s)-encoding from $\mathbb{F}_p^{\ell}$ to $\mathbb{F}_p^{\bar{\ell}}$.

Consider the following setting: Let $\mathbf{Z} = (\mathbf{X} \circ \mathbf{Y})_t$ be any 2-interleaved source on $\{0,1\}^{2n}$, where $\mathbf{X}$ and $\mathbf{Y}$ are arbitrary independent sources on $\{0,1\}^n$ with min-entropy k_1 and k_2 respectively, and $t : [2n] \to [2n]$ is any permutation.

Our first step is to use an (n,s)-encoding enc from $\mathbb{F}_2^{2n}$ to $\mathbb{F}_2^{\bar{n}}$ to encode $\mathbf{Z}$. Thus,

$$\mathrm{enc}(\mathbf{Z}) = \mathbf{X}' + \mathbf{Y}',$$

where

$$\mathbf{X}' = \sum_{i=1}^{n} \mathbf{X}_i v_{t(i)}, \quad \mathbf{Y}' = \sum_{i=j}^{n} \mathbf{Y}_j v_{t(n+j)}.$$

where $S = \{v_1, \ldots, v_{2n}\}$ is an (n,s)-spanning set of vectors.

The idea is to argue that the independent sources $\mathbf{X}'$ and $\mathbf{Y}'$ (on $\{0,1\}^{\bar{n}}$) have enough min-entropy. Since (by construction) the span of the set of vectors $\{v_{t(1)}, \ldots, v_{t(n)}\}$ has dimension at least s, Lemma 2.3.12 implies that $H_\infty(\mathbf{X}') = k_1' \geq k_1 - (n-s)$. Similarly $H_\infty(\mathbf{Y}') = k_2' \geq k_2 - (n-s)$.

We now associate $\mathbb{F}_2^{\bar{n}}$ with $\mathbb{F}_{2^{\bar{n}}}$. A character sum estimate of Karatsuba[2] [Kar71, Kar91] implies that for any nonprincipal multiplicative character χ of $\mathbb{F}_{2^{\bar{n}}}^*$,

$$\mathbf{E}_{\mathbf{X}'} |\mathbf{E}_{\mathbf{Y}'} [\chi(\mathbf{X}' + \mathbf{Y}')]| \leq 2^{-\delta k_2'}$$

whenever: $k_1 \geq (\frac{1}{2} + 3\delta)\bar{n} + (n-s)$ and $k_2 \geq \frac{4}{\delta} \log \bar{n} \log p + (n-s)$.

[2] this character sum was also used in [CG88] for constructing explicit two-source extractors.

Suppose k_1 and k_2 satisfy these conditions.

We then follow a standard approach and define the function:

$$\text{ext}(\mathbf{Z}) = \log_g(\mathbf{X}' + \mathbf{Y}') \pmod{M},$$

where $M = 2^{\delta k_2'/2}$ and g is a primitive element of $\mathbb{F}_{2^{\bar{n}}}$. Using a version of the Abelian XOR lemma (see Lemma 2.6.2), it follows that ext is an extractor with output length $\delta k_2'/2$ and error $2^{-\Omega(k_2')}$. Further the extractor is strong in the source $\mathbf{X}$. However, the running time of this extractor is subexponential since it involves computing discrete logs over finite fields. This gives us a semi-explicit extractor construction.

To get a polynomial time extractor, we compute discrete log over a smaller multiplicative subgroup of $\mathbb{F}^*_{2^{\bar{n}}}$. Let $M|2^{\bar{n}} - 1$ and $M = n^{\lambda}$ for any constant λ (we show in Theorem 10.4.2 that we can ensure that there is always such an M). Define the function:

$$\text{ext}_1(\mathbf{Z}) = \text{enc}(\mathbf{Z})^{\frac{2^{\bar{n}}-1}{M}}.$$

Thus $\text{ext}_1(\mathbf{Z})$ is a distribution on the multiplicative subgroup $G = \{x^{\frac{2^{\bar{n}}-1}{M}} : x \in \mathbb{F}^*_{2^{\bar{n}}}\}$ (of $\mathbb{F}^*_{2^{\bar{n}}}$) of size M (in fact $\text{ext}_1(\mathbf{Z})$ is a distribution on $G \cup \{0\}$, but $\Pr[\text{ext}_1(\mathbf{Z}) = 0] = 2^{-\Omega(n)}$ and hence we ignore this and add this to the error). Let g be a generator of G. It now follows by using the character sum estimate of Karatsuba [Kar71] that the function:

$$\text{ext}(\mathbf{Z}) = \log_g(\text{ext}_1(\mathbf{Z}))$$

is an extractor.

We need to find a generator g of G efficiently. For this, we use an efficient algorithm of Shoup [Sho90] for finding a small set of elements such that one of them is a primitive element of $\mathbb{F}_{2^{\bar{n}}}$. We use a straightforward method to find g from this set in polynomial time. We achieve output length of $\lambda \log n$ and error $n^{-\Omega(1)}$. The extractor is strong in the source $\mathbf{X}$.

Reducing the Min-Entropy Rate For some c and any $\delta > 0$, let $p > 2^{\frac{c}{\delta}}$ be any prime. When the source $\mathbf{Z} = (\mathbf{X} \circ \mathbf{Y})_t$ is on $[p]^{2n}$, we can reduce the min-entropy rate requirement of the source $\mathbf{X}$ to $(\frac{1}{2} + \delta)$. The construction follows the same outline as above (using (n, s)-encodings from $\mathbb{F}_p^{2n}$ to $\mathbb{F}_p^{\bar{n}}$), and the improvement is achieved by using the fact that over alphabet $[p]$, we can construct (n, n)-spanning sets in $\mathbb{F}_p^{\bar{n}}$ with $\bar{n} = n(1 + \frac{\delta}{5})$ (using explicit codes from [GI02]). The output length of the extractor obtained is $\lambda \log n$ (for any constant λ) and achieves error $n^{-\Omega(1)}$. Further the extractor is strong in the source $\mathbf{X}$.

Improving the Output Length We improve the output length of the above extractor to $\Omega(n)$ when both sources $\mathbf{X}$ and $\mathbf{Y}$ (on $[p]^n$) have min-entropy at least $(\frac{1}{2} + \delta)n \log p$. Our construction is as follows. Let SExt be an explicit strong seeded extractor for linear min-entropy with linear output length and polynomially small error with seed seed length $O(\log n)$, for example from the work of [GUV09]. Let $\mathbf{Z}_{[n]}$ denote the projection of $\mathbf{Z}$ to the first n coordinates and let ext_p denote the extractor constructed in the previous paragraph (for 2-interleaved sources on $[p]^{2n}$). Our extractor is the following function:

$$\mathrm{ext}_{p,long}(\mathbf{Z}) = \mathrm{SExt}(\mathbf{Z}_{[n]}, \mathrm{ext}_p(\mathbf{Z})).$$

We sketch the proof of correctness. Without loss of generality, suppose that $\mathbf{X}$ has more symbols in $\mathbf{Z}_{[n]}$ than the source $\mathbf{Y}$. Let $S \subseteq [n]$ be the coordinates of $\mathbf{X}$ which are in $\mathbf{Z}_{[n]}$ and let $\mathbf{X}_S$ denote the projection of $\mathbf{X}$ to the coordinates indexed by S. Let $T \subset [n]$ be the coordinates of $\mathbf{Y}$ which are in $\mathbf{Z}_{[n]}$ and let $\mathbf{Y}_T$ denote the projection of $\mathbf{Y}$ to the coordinates indexed by T. Further, we use $\mathbf{X}_S \circ \mathbf{Y}_T$ to denote $\mathbf{Z}_{[n]}$. Note that, by assumption $|S| \geq \frac{n}{2}$ and $|T| \leq \frac{n}{2}$. It follows by Lemma 2.3.7 that $\mathbf{Y}|\mathbf{Y}_T$ is close to a source with min-entropy $> \frac{\delta n \log p}{2}$ with probability $1 - 2^{-\Omega(n)}$. Also note that $\mathbf{X}_S$ has min-entropy $\geq \delta n \log p$.

Consider such a good fixing $\mathbf{Y}_T = y_T$. Since $\mathbf{X}$ and $\mathbf{Y}|\mathbf{Y}_T = y_T$ have enough min-entropy, it follows that even under this fixing, $W = \mathrm{ext}_p(\mathbf{Z})$ is close to uniform. We now use the property that ext_p is strong with respect to the source $\mathbf{X}_S$, i.e.,

$$|(\mathbf{X}_S, W) - (\mathbf{X}_S, U_d)| \leq n^{-\Omega(1)}.$$

Using a probability lemma from [Sha06], it follows that for any $W = w$,

$$|\mathbf{X}_S - (\mathbf{X}_S|(W = w))| \leq n^{-\Omega(1)},$$

(using that w is of length $O(\log n)$).

Hence, $\mathrm{SExt}(\mathbf{X}_S \circ \mathbf{Y}_T, W)|\mathbf{Y}_T = y_T$ is $n^{-\Omega(1)}$-close to the convex combination: $\sum_w \Pr[(W|\mathbf{Y}_T = y_T) = w]\mathrm{SExt}(\mathbf{X}_S \circ \mathbf{Y}_T, w)|\mathbf{Y}_T = y_T$. Since as observed above, $W|\mathbf{Y}_T = y_T$ is $n^{-\Omega(1)}$-close to U_d, it follows that $\mathrm{SExt}(\mathbf{X}_S \circ \mathbf{Y}_T, W)|\mathbf{Y}_T = y_T$ is $n^{-\Omega(1)}$-close to $\mathrm{SExt}(\mathbf{X}_S \circ y_T, U_d)$. The correctness now follows using the fact that SExt is a seeded extractor for linear min-entropy.

Probabilistic Method We show in Lemma 10.3.10, that a random set $S \subset \mathbb{F}_2^n$ of size $2n$ is an $(n, n - 2\sqrt{n})$-spanning set with high probability. Thus, using the proof technique described above, any explicit construction of such a set will yield explicit extractors for 2-interleaved sources on $\{01\}^{2n}$ when one source has min-entropy at least $0.51n$ and the other source has min-entropy at least $cn^{\frac{1}{2}}$. We leave it as an interesting open problem to explicitly construct such a set S.[3]

We give formal proofs of the above extractor constructions and other related constructions in Section 10.4.

10.2.2 Interleaved Non-Malleable Extractors

For some $c > 0$ and any $\delta > 0$, let $p > 2^{\frac{c}{\delta}}$ be any prime. Let $\mathbf{X}$ be a source on $[p]^n$ with min-entropy k_1 and $\mathbf{Y}$ be a weak- eed on $[p]^n$ with min-entropy k_2. Let $f : [p]^n \to [p]^n$ be any function with no fixed points. Thus the non-malleable extractor has access to $\mathbf{Z} = (\mathbf{X} \circ \mathbf{Y})_t$ for an artitrary permutation $t : [2n] \to [2n]$. Let $\mathbf{Z}_f$ denote the tampered source $(\mathbf{X} \circ f(\mathbf{Y}))_t$.

We show that the extractor ext_p constructed for 2-interleaved sources (described in the previous section) is also non-malleable. We prove it in the following way. Recall the construction

[3]This is related to finding explicit constructions of binary erasure list-decodable codes with almost optimal parameters. See Section 10.3 for more details.

of ext_p:

$$\mathrm{enc}(\mathbf{Z}) = \sum_{i=1}^{2n} \mathbf{Z}_i v_i, \quad \mathrm{ext}_1(\mathbf{Z}) = \mathrm{enc}(\mathbf{Z})^{\frac{p^{\bar{n}}-1}{M}}, \quad \mathrm{ext}_p(\mathbf{Z}) = \log_g(\mathrm{ext}_1(\mathbf{Z})),$$

where $S = \{v_1, \ldots, v_{2n}\}$ is an (n, n)-spanning set in $\mathbb{F}_p^{\bar{n}}$, $M = \mathrm{poly}(n)$, $\bar{n} = n(1 + \frac{\delta}{5})$ and g is a generator of the multiplicative subgroup $G = \{x^{\frac{2^{\bar{n}}-1}{M}} : x \in \mathbb{F}_{2^{\bar{n}}}^*\}$.

Since ext_p is a distribution on $\mathbf{Z}_M$, it follows by a version of the Abelian XOR lemma proved in [DLWZ14] that to prove non-malleability, it is enough to prove the bound:

$$|\mathbf{E}[\psi_a(\mathrm{ext}_p(\mathbf{Z}))\psi_b(\mathrm{ext}_p(\mathbf{Z}_f))]| \leq n^{-\Omega(1)},$$

for all additive characters ψ_a and ψ_b (of $\mathbf{Z}_M$) such that ψ_a is nontrivial. When ψ_b is the trivial character, the above quantity can be bounded by the fact that ext_p is an extractor for 2-interleaved sources. Thus, suppose both ψ_a and ψ_b are nontrivial.

It follows that

$$|\mathbf{E}[\psi_a(\mathrm{ext}_p(\mathbf{Z}))\psi_b(\mathrm{ext}_p(\mathbf{Z}_f))]| = |\mathbf{E}[\chi_a(\mathrm{enc}(\mathbf{Z}))\chi_b(\mathrm{enc}(\mathbf{Z}_f))]|$$

where χ_a and χ_b are nonprincipal multiplicative characters of $\mathbb{F}_{2^{\bar{n}}}^*$.

Further, $\mathbf{Z} = \sum_{i=1}^n \mathbf{X}_i v_{t(i)} + \sum_{j=1}^n \mathbf{Y}_j v_{t(j)}$ and $\mathbf{Z}_f = \sum_{i=1}^n \mathbf{X}_i v_{t(i)} + \sum_{j=1}^n f(\mathbf{Y})_j v_{t(j)}$. Thus,

$$\mathbf{Z} = \mathbf{X}' + \mathbf{Y}', \quad \mathbf{Z}_f = \mathbf{X}' + f'(\mathbf{Y}'),$$

where $\mathbf{X}' = \sum_{i=1}^n \mathbf{X}_i v_{t(i)}$, $\mathbf{Y}' = \sum_{i=j}^n \mathbf{Y}_j v_{t(n+j)}$ and $f' = L \circ f \circ L^{-1}$, L being the one-one linear map $L(z) = \sum_{i=1}^n z_i v_{t(n+i)}$. Thus,

$$|\mathbf{E}[\psi_a(\mathrm{ext}_p(\mathbf{Z}))\psi_b(\mathrm{ext}_p(\mathbf{Z}_f))]| = |\mathbf{E}[\chi_a(\mathbf{X}' + \mathbf{Y}')\chi_b(\mathbf{X}' + f'(\mathbf{Y}'))]|.$$

Using the work of Dodis et al. [DLWZ14], we can prove the required upper bound on the quantity on the right hand side if f' does not have any fixed points. We indeed show that f' has no fixed

points (by using the fact that L is one-one and f has no fixed points). This completes the proof sketch. The non-malleable extractor outputs $\lambda \log n$ bits (for any constant λ) and achieves error $n^{-\Omega(1)}$. See Section 10.5 for more details.

Notation

For any permutation $t : [n] \to [n]$, define the string $w = (s)_t \in [R]^n$ such that $w_i = s_{t(i)}$ for $i = 1, \ldots, n$. Further for any $t \subset [n]$, let s_T denote the $|T|$ length string that is the projection of s onto the coordinates indexed by T.
For any $x \in [p]^{n_1}$, $y \in [p]^{n_2}$ and disjoint subsets $S, T \subset [n_1 + n_2]$ with $|S| = n_1$, $|T| = n_2$, we define $z = x_S \circ y_T$ such that $z_S = x$ and $z_T = y$.

10.3 Constructing Spanning Vectors

A key ingredient in our extractor construction are explicit constructions of spanning vectors. Recall that a set of vectors $S \subseteq \mathbb{F}_p^{\bar{\ell}}$ is (r, s)-spanning if the span of any r vectors of S has dimension at least s (see Definition 10.2.1). Our constructions of spanning vectors are simple and are based on explicit linear codes. Recall that a linear code of block length n, dimension k and distance d over any field $\mathbb{F}$ is a k dimensional subspace over $\mathbb{F}$ with the number of zero coordinates of any vector in this subspace being at most $n - d$. The relative rate of the code is k/n and the relative distance is d/n.

We show that the columns of the parity check matrix of any linear code with good erasure list-decoding radius (defined below) can be used as a spanning set.

Definition 10.3.1 (Erasure List-Decoding Radius [Gur03])**.** *We say that a linear code $[n, k, d]$ code $\mathcal{C}$ over a finite field $\mathbb{F}$ is (e, L)-erasure list-decodable if for every for every $r \in \mathbb{F}^{n-e}$ and $T \subseteq [n]$ of size $n - e$, $|\{c \in \mathcal{C} : c_T = r\}| \leq L$.*

We now establish a simple connection between erasure list-decodable codes and spanning

sets.

Lemma 10.3.2. *Let $\mathcal{C}$ be a linear $[n, k, d]$ code over a finite field $\mathbb{F}$, which is (e, L)-earasure list-decodable. Let H be parity check matrix of $\mathcal{C}$, and let S be the set of columns of H. Then $S \subset \mathbb{F}^{n-k}$ is a (r, s)-spanning set of size n, with $r = e$ and $s = e - \log_{|\mathbb{F}|}(L)$.*

Proof. Since $\mathcal{C}$ is (e, L)-erasure list-decodable, it follows that the size of the null space of any e columns of the parity check matrix H is at most L. By the rank-nullity theorem, it follows that the rank of the sub-matrix of H restricted to these e columns is at least $e - \log_{|\mathbb{F}|}(L)$. Thus by definition, the set of columns of H form a $(e, e - \log_{|\mathbb{F}|}(L))$-spanning set. □

The following lemma relates the minimum distance of a code to its erasure list-decoding radius, and can be seen as an analogue of the Johnson bound for erasure list-decoding.

Lemma 10.3.3 ([Gur04b])**.** *Let $\mathcal{C}$ be a code with block length n and relative distance δ over an alphabet of size q. Then for any $\epsilon > 0$, $\mathcal{C}$ is a (e, L)-erasure list-decodable code, where $e = \left(\frac{q}{q-1} - \epsilon\right)\delta n$ and $L = \frac{q}{(q-1)\epsilon}$.*

Combining the above results, the following lemma is immediate.

Lemma 10.3.4. *For any $\delta > 0$, let $\mathcal{C}$ be a binary linear code with relative distance $\frac{1}{4} + \delta$, and block length $2n$. Then the columns of the parity check matrix of H form a (r, s)-spanning set, with $r = n$ and $s = n - \log\left(\frac{1}{\delta}\right)$.*

Proof. Using Lemma 10.3.3, it follows that $\mathcal{C}$ is $(n, \frac{1}{\delta})$-erasure list-decodable. Now applying Lemma 10.3.2, the lemma follows directly. □

A similar result follows for the case of q-ary linear codes.

Lemma 10.3.5. *For any $\delta > 0$, let $\mathcal{C}$ be a linear code with relative distance $\frac{q-1}{2q} + \delta$ and block length $2n$ over a finite field of size q. Then the columns of the parity check matrix of H form a (r, s)-spanning set, with $r = n$ and $s = n - \log\left(\frac{q}{(q-1)\delta}\right)$.*

To instantiate the above results, we recall some explicit code constructions. Using standard code concatenation, there are known constructions of binary linear codes achieving the Zyablov bound.

Theorem 10.3.6. *For any $\epsilon, \gamma > 0$, there exists an explicit construction of a binary linear code with relative distance $\delta = \frac{1}{4} + \epsilon$ and relative rate $R \geq \max_{0<r<1-H(\delta+\epsilon)} r\left(1 - \frac{\delta}{H^{-1}(1-r)-\epsilon}\right)$.*

Over larger alphabets, the following explicit codes were constructed in the work of Guruswami and Indyk [GI02].

Theorem 10.3.7 ([GI02])**.** *There exists $c > 0$ such that for every $\gamma > 0$ and any prime $p > 2^{\frac{c}{\gamma}}$ there is an efficient construction of a linear code $C \subset \mathbb{F}_p^n$ with relative distance $\delta = \frac{1}{2} - \frac{1}{4p}$ and rate $R = \frac{1}{2} - \gamma$.*

Using the above codes, we now have explicit constructions of spanning sets.

Lemma 10.3.8. *There exist constants $\gamma > 0$ and c such that for any n, there exists an explicit $(n, n-c)$-spanning set $S \subset \mathbb{F}_{2^{\bar{n}}}$ of size $2n$, where $\bar{n} = 2n(1-\gamma)$.*

Proof. Let H be the parity check matrix of the explicit linear code $C \subset \mathbb{F}_2^{2n}$ from Theorem 10.3.6 for relative distance $\frac{1}{4} + \delta$, for some small constant δ. Let $S = \{v_1, \ldots, v_{2n}\}$ be the set of columns of H. Thus $S \subset \mathbb{F}_2^{\bar{n}}$, $\bar{n} = 2n(1-\gamma)$, γ being the relative rate of the code. Applying Lemma 10.3.4, the result is now immediate. □

Lemma 10.3.9. *There exists $c > 0$ such that for any $\gamma > 0$ and any prime $p > 2^{\frac{c}{\gamma}}$, there is an efficient construction of an explicit set $(n, n-C)$-spanning set $S \subset \mathbb{F}_{2^{\bar{n}}}$ of size $2n$, where $\bar{n} = n(1+2\gamma)$ and $C = \frac{2c}{\gamma}$.*

Proof. Let H be the parity check matrix of the explicit linear code $C \subset \mathbb{F}_p^{2n}$ from Theorem 10.3.7 with relative distance $\frac{1}{2} - \frac{1}{4p}$ and rate $\frac{1}{2} - \gamma$. Let $S = \{v_1, \ldots, v_{2n}\}$ be the set of columns of H. The result now follows by Lemma 10.3.5. □

We show that random sets are (r, s)-spanning sets with overwhelmingly high probability. Guruswami's existence proof of subspace evasive [Gur11] targets different parameters and does not apply here. This lemma is more related to the existence of good erasure list-decodable codes.

Lemma 10.3.10. *Let S be a random subset of $\mathbb{F}_2^n$ of size $2n$. Then,*

$$\Pr[S \text{ is not a } (n, n-2\sqrt{n})\text{-spanning set }] \leq 2^{-n}.$$

Proof. Let $t > 0$. Consider any subset $R \subset S$, $|R| = n$. By standard arguments, it follows that

$$\Pr[\dim(\mathrm{span}(R)) \leq n - t] \leq \binom{n}{t}(2^{-t})^t \leq \left(\frac{n}{2^t}\right)^t.$$

Thus,

$$\Pr[\exists\ R \subset S,\ |R| = n \text{ with } \dim(\mathrm{span}(R)) \leq n - t] \leq \binom{2n}{n}\left(\frac{n}{2^t}\right)^t \leq 2^{2n - t^2 + t\log n}$$

The lemma follows by setting $t = 2\sqrt{n} + 1$. □

10.4 Extractors for 2-Interleaved Sources

10.4.1 Extractors for 2-Interleaved Sources on $\{0,1\}^{2n}$

Our extractor constructions are based on encoding the interleaved-sources using spanning vectors. Recall that any (r, s)-encoding from $\mathbb{F}_p^\ell \to \mathbb{F}_p^{\bar{\ell}}$ is defined in the following way: For any (r, s)-spanning set $S = \{v_1, \ldots, v_\ell\} \subseteq \mathbb{F}_p^{\bar{n}}$, the function $\mathrm{enc} : \mathbb{F}_p^\ell \to \mathbb{F}_p^{\bar{\ell}}$ defined as

$$\mathrm{enc}(z) = \sum_{i=1}^{n} z_i v_i$$

is an (r, s)-encoding from $\mathbb{F}_p^\ell \to \mathbb{F}_p^{\bar{\ell}}$.

The following is a key lemma in our extractor constructions.

Lemma 10.4.1 (Main Lemma)**.** *Fix any $\delta > 0$. Let p be any prime and let $\mathbf{Z} = (\mathbf{X} \circ \mathbf{Y})_t$ be any 2-interleaved source on $\mathbb{F}_p^{2n}$, where $\mathbf{X}$ and $\mathbf{Y}$ are independent sources on $\mathbb{F}_p^n$ with min-entropy k_1 and k_2 respectively, and $t : [2n] \to [2n]$ is any permutation. Also suppose χ is any nonprincipal multiplicative character of $\mathbb{F}_{p^{\bar{n}}}^*$ and* enc *is an arbitrary (n, s)-encoding from $\mathbb{F}_p^{2n}$ to $\mathbb{F}_p^{\bar{n}}$. Then,*

$$\mathbf{E}_{\mathbf{X}}|\mathbf{E}_{\mathbf{Y}}[\chi(\mathrm{enc}(\mathbf{Z}))]| \leq 2^{-\delta(k_2-(n-s)\log p)},$$

whenever

- $k_1 \geq (\frac{1}{2} + 3\delta)\bar{n}\log p + (n-s)\log p$*, and*
- $k_2 \geq \frac{4\log \bar{n}\log p}{\delta} + (n-s)\log p$*.*

Proof. For any $z \in \mathbb{F}_p^{2n}$, let

$$\mathrm{enc}(z) = \sum_{i=1}^{2n} z_i v_i$$

where $S = \{v_1, \ldots, v_{2n}\} \subset \mathbb{F}_p^{\bar{n}}$ is (n, s)-spanning.

We have,

$$\chi(\mathrm{enc}(\mathbf{Z})) = \chi\left(\sum_{i=1}^{2n} \mathbf{Z}_i v_i\right) = \chi\left(\sum_{i=1}^{n} \mathbf{X}_i v_{t(i)} + \sum_{j=1}^{n} \mathbf{Y}_j v_{t(n+j)}\right)$$

Define the following independent sources:

$$\mathbf{X}' = \sum_{i=1}^{n} x_i v_{t(i)} : x \sim \mathbf{X}, \qquad \mathbf{Y}' = \sum_{j=1}^{n} y_j v_{t(n+j)} : y \sim \mathbf{Y}.$$

Using Lemma 2.3.12, it follows that: $k_1' = H_\infty(\mathbf{X}') \geq k_1 - (n-s)\log p$ and $k_2' = H_\infty(\mathbf{Y}') \geq k_2 - (n-s)\log p$.

Thus, we have

$$\begin{aligned}\mathbf{E}_{\mathbf{X}}|\mathbf{E}_{\mathbf{Y}}[\chi(\mathrm{enc}(\mathbf{Z}))]| &= \mathbf{E}_{x\sim\mathbf{X}}\left|\mathbf{E}_{y\sim\mathbf{Y}}\left[\chi\left(\sum_{i=1}^{n} x_i v_{t(i)} + \sum_{j=1}^{n} y_j v_{t(n+j)}\right)\right]\right| \\ &= \mathbf{E}_{\mathbf{X}'}\left|\mathbf{E}_{\mathbf{Y}'}\left[\chi\left(\mathbf{X}'+\mathbf{Y}'\right)\right]\right| \\ &= 2^{-\delta k_2'}\end{aligned}$$

where the last inequality follows using Theorem 2.5.5. □

Using the above main lemma, we construct extractors for 2-interleaved sources on $\mathbb{F}_2^{2n}$.

Theorem 10.4.2. *For some $\delta > 0$ and any $\lambda > 0$, there exists an explicit function* $\mathrm{ext} : \{0,1\}^{2n} \to [M]$, *$M = n^{\lambda}$, such that if $\mathbf{X}$ and $\mathbf{Y}$ are independent sources on $\mathbb{F}_2^n$ with min-entropy k_1, k_2 respectively satisfying $k_1 > (1-\delta)n$ and $k_2 > 35\max\{\log n, \log M\}$, $t : [2n] \to [2n]$ is any permutation, then*

$$|\mathrm{ext}((\mathbf{X}\circ\mathbf{Y})_t)\circ\mathbf{X} - U_M \circ \mathbf{X}| = 2^{-\Omega(k_2)}.$$

Proof. Let H be the parity check matrix of a code $C \subset \mathbb{F}_2^{2n}$ with relative distance $= \frac{1}{4} + \delta_1$ (for some small constant δ_1) and constant rate R, where we fix R as follows. Let R_Z be the rate of the code from Theorem 10.3.6. Let $\epsilon_1 << R_Z$ be a small constant. We choose R in the interval $[R_Z - \epsilon_1, R_Z]$ such that $\bar{n} = 2n(1-R)$ is divisible by integer m, $m = \lambda \log n$. Since $2R_Z\epsilon_1 n >> m$, we can indeed find such an R. Fix $M = 2^m - 1$. We note that $M|2^{\bar{n}} - 1$. Set $\delta = \frac{R}{6}$.

Let $S = \{v_1, \ldots, v_{2n}\}$ be the set columns of H. By Lemma 10.3.8, S is $(n, n-C)$-spanning, for some constant C. We interpret each v_i as being an element in the field $\mathbb{F}_{2^{\bar{n}}}$. Consider the multiplicative subgroup:

$$G = \{x^{\frac{2^{\bar{n}}-1}{M}} : x \in \mathbb{F}_{2^{\bar{n}}}^*\}.$$

A generator g of G can be found efficiently in the following way: Using Theorem 2.7.1, we can efficiently construct a set $S = \{a_1, \ldots, a_l\}$, $l = \mathrm{poly}(n)$, such that one of the a_i's, say a_j, is a

primitive element of $\mathbb{F}_{2^{\bar{n}}}$. Let $S' = \{a_1^{\frac{2^{\bar{n}}-1}{M}}, \ldots, a_l^{\frac{2^{\bar{n}}-1}{M}}\}$. We note that $a_j^{\frac{2^{\bar{n}}-1}{M}} \in S'$ is an element of order M. Thus, it is enough to enumerate over the elements in S' and compute the order of each element. Since the order of any element in S' is bounded by $M = \text{poly}(n)$, the search procedure can be implemented efficiently.

Let $\mathbf{Z} = (\mathbf{X} \circ \mathbf{Y})_t$. For any $z \in \mathbb{F}_2^{2n}$, define the functions:

$$\text{enc}(z) = \sum_{i=1}^{2n} z_i v_i, \quad \text{ext}_1(z) = (\text{enc}(z))^{\frac{2^{\bar{n}}-1}{M}}, \quad \text{ext}(z) = \log_g(\text{ext}_1(z)).$$

We note that ext_1 and ext are efficiently computable functions. Further note that enc is an $(n, n-C)$-encoding from $\mathbb{F}_2^{2n}$ to $\mathbb{F}_2^{\bar{n}}$.

Using the above lemma, we prove the following claim.

Claim 10.4.3. *Let $\psi(x) = e_M(\beta x)$, $\beta \neq 0 \pmod M$, be any nontrivial character of the additive group $\mathbf{Z}_M$.*

Then,

$$\mathbf{E}_{\mathbf{X}} \left| \mathbf{E}_{\mathbf{Y}} [\psi(\text{ext}_2((\mathbf{X} \circ \mathbf{Y})_t))] \right| \leq 2^{-\delta k_2}.$$

We note that Theorem 10.4.2 follows directly from Claim 10.4.3 by using Lemma 2.6.1. Thus it is enough to prove Claim 10.4.3.

Proof of Claim 10.4.3. We have,

$$\begin{aligned} \psi(\text{ext}(z)) &= e_M(\beta \log_g(\text{ext}_1(z))) \\ &= \chi\left(\text{enc}(z)\right), \end{aligned}$$

where $\chi(x) = e_M(\beta \log_g(x))$ is a nonprincipal multiplicative character of $\mathbb{F}_{2^{\bar{n}}}^*$ of order $\frac{M}{\gcd(M,\beta)}$.

Thus, we have

$$\begin{aligned}\mathbf{E}_{\mathbf{X}}\left|\mathbf{E}_{\mathbf{Y}}[\psi(\mathrm{ext}_2((\mathbf{X}\circ\mathbf{Y})_t))]\right| &= \mathbf{E}_{x\sim\mathbf{X}}\left|\mathbf{E}_{y\sim\mathbf{Y}}\left[\chi\left(\mathrm{enc}(\mathbf{Z})\right)\right]\right| \\ &\leq 2^{-\delta k_2},\end{aligned}$$

where the inequality follows from Lemma 10.4.1. □

□

It is direct from the above theorem, that if we insist that the output of the above extractor is a bit string, we have the following result.

Theorem 10.4.4 (Theorem 23 restated)**.** *For some $\delta > 0$ and any $\lambda > 0$, there exists an explicit function $\mathrm{ext} : \{0,1\}^{2n} \to \{0,1\}^m$, $m = \lambda \log n$, such that if $\mathbf{X}$, $\mathbf{Y}$ are independent sources on $\mathbb{F}_2^n$ with min-entropy k_1, k_2 respectively satisfying $k_1 > (1-\delta)n$ and $k_2 > 35\max\{\log n, m\}$, $t : [2n] \to [2n]$ is any permutation, then*

$$|\mathrm{ext}((\mathbf{X}\circ\mathbf{Y})_t)\circ\mathbf{X} - U_m \circ \mathbf{X}| = n^{-\Omega(1)}.$$

10.4.2 Extracting from 2-Interleaved Sources on $\mathbb{F}_p^{2n}$

If the sources $\mathbf{X}$ and $\mathbf{Y}$ are on $\mathbb{F}_p^n$ (for some large enough prime p), we can reduce the min-entropy rate requirement of the source $\mathbf{X}$ to about $\frac{1}{2}$.

Theorem 10.4.5 (Theorem 24 restated)**.** *There exists $c > 0$ such that for any $\delta, \lambda > 0$ and any prime $p > 2^{\frac{c}{\delta}}$, there exists an explicit function $\mathrm{ext}_p : \mathbb{F}_p^{2n} \to \{0,1\}^m$, $m = \lambda \log n$, such that if $\mathbf{X}$ and $\mathbf{Y}$ are independent sources on $\mathbb{F}_p^n$ with min-entropy k_1, k_2 respectively, satisfying $k_1 > (\frac{1}{2}+\delta)n\log p$ and $k_2 > \frac{5}{\delta}\max\{\log n \log p, m\}$, $t : [2n] \to [2n]$ is any injective map, then*

$$|\mathrm{ext}_p((\mathbf{X}\circ\mathbf{Y})_t)\circ\mathbf{X} - U_m \circ \mathbf{X}| = n^{-\Omega(1)}.$$

Proof. Let $S = \{v_1, \ldots, v_{2n}\}$ be an explicit (n,n-C)-spanning set in $\mathbb{F}_p^{\bar{n}}$ from Lemma 10.3.9. Further, as in the proof of Theorem 10.4.2, we choose the rate of the code in Lemma 10.3.9 such that $m|\bar{n}$ and $m = \lambda \log_p n$. Thus we can ensure that $\bar{n} \leq n(1 + \frac{\delta}{5})$.

Let $M = n^\lambda$. For any $z \in \mathbb{F}_p^{2n}$, define the functions:

$$\mathrm{enc}(z) = \sum_{i=1}^{2n} z_i v_i, \quad \mathrm{ext}_1(z) = (\mathrm{enc}(z))^{\frac{p^{\bar{n}}-1}{M}}, \quad \mathrm{ext}(z) = \log_g(\mathrm{ext}_1(z))$$

where g is a generator of $G = \{x^{\frac{p^{\bar{n}}-1}{M}} : x \in \mathbb{F}_{p^{\bar{n}}}^*\}$. The proof now follows using Lemma 10.4.1 and Lemma 2.6.1. □

10.4.3 Improving the Output Length

The output length of the extractor in Theorem 10.4.5 is $\Omega(\log n)$. We improve the output length to $\Omega(n)$ bits when the min-entropy rate of both the sources (on $\mathbb{F}_p^n$) are slightly more than $\frac{1}{2}$.

A general technique to improve the output length extractors was introduced by Shaltiel [Sha06]. In particular, Shaltiel showed that the function:

$$\mathrm{SExt}(\mathbf{X}, 2\mathrm{ext}(\mathbf{X}, \mathbf{Y})) \circ \mathrm{SExt}(\mathbf{Y}, 2\mathrm{ext}(\mathbf{X}, \mathbf{Y}))$$

is 2-source extractor with longer output length, where 2ext is a 2-source extractor with short output length and SExt is a seeded extractor set to appropriate parameters.

However this does not work in our case since it requires access to the individual sources $\mathbf{X}$ and $\mathbf{Y}$. Surprisingly, we show that the construction: $\mathrm{SExt}(((\mathbf{X} \circ \mathbf{Y})_t)_{[n]}, 2\mathrm{ext}_p((\mathbf{X} \circ \mathbf{Y})_t))$ can be proved to be an extractor.

Theorem 10.4.6. *There exists $c > 0$ such that for any $\delta > 0$ and any prime $p > 2^{\frac{c}{\delta}}$, there exists an explicit function $\mathrm{ext}_{p,long} : \mathbb{F}_p^{2n} \to \{0,1\}^m$, $m = \Omega(n)$, such that if $\mathbf{X}$ and $\mathbf{Y}$ are independent sources on $\mathbb{F}_p^n$ with min-entropy k_1, k_2 respectively satisfying $k_1 > (\frac{1}{2} + \delta)n \log p$ and $k_2 > (\frac{1}{2} + \delta)n \log p$,*

$t : [2n] \to [2n]$ *is any injective map, then*

$$|\mathrm{ext}_{p,long}((\mathbf{X} \circ \mathbf{Y})_t) - U_m| = n^{-\Omega(1)}.$$

Proof. Let SExt be the seeded-extractor from Theorem 2.1.2 with parameters $\beta = \delta$, $\alpha = \delta/2$ and $\epsilon = n^{-\Omega(1)}$. Let the seed length of SExt with this setting of the parameters be $d = \lambda \log n$. Let $\mathbf{Z} = (\mathbf{X} \circ \mathbf{Y})_t$. Define

$$\mathrm{ext}_{p,long}(\mathbf{Z}) = \mathrm{SExt}(\mathbf{Z}_{[n]}, \mathrm{ext}_p(\mathbf{Z})),$$

where ext_p is the extractor from Theorem 10.4.5 designed to extract from 2-interleaved sources with one source at min-entropy $k_1 \geq (\frac{1}{2} + \delta)n \log p$ and the other source with min-entropy $k_2 \geq \frac{\delta n \log p}{2}$ with error $\epsilon_p = n^{-2\lambda}$ and output length $m_p = \lambda \log n$.

Let $S = \{i \in [n] : \mathbf{Z}_i = \mathbf{X}_i\}$ and $T = \{j \in [n] : \mathbf{Z}_j = \mathbf{Y}_j\}$. Also let $\bar{S} = [n] \setminus S$ and $\bar{T} = [n] \setminus T$. Without loss of generality, we can assume that $|S| \geq \frac{n}{2}$. It follows from Lemma 2.3.7 that there exists a set Good_y such that for any $y_T \in \mathit{Good}_y$, $\mathbf{Y}_{\bar{T}} | \mathbf{Y}_T = y_T$ is $2^{-\Omega(n)}$-close to a source with entropy more than $\frac{\delta n \log p}{2}$, and $\Pr[\mathbf{Y}_t \in \mathit{Good}_y] > 1 - 2^{-\Omega(n)}$.

Let $y_T \in \mathit{Good}_y$. It follows by the setting of ext_p that

$$|(\mathrm{ext}_p(\mathbf{Z}|\mathbf{Y}_T = y_T) \circ \mathbf{X}_S - U_m \circ \mathbf{X}_S| \leq n^{-2\lambda}.$$

Using Lemma 2.3.9, it follows that

$$|\mathbf{X}_S - (\mathbf{X}_S | (\mathrm{ext}_p(\mathbf{Z}|\mathbf{Y}_T = y_T) = e))| \leq n^{-\lambda+1}. \tag{10.1}$$

Let $p_{y_T} = \Pr[\mathbf{Y}_T = y_T]$ and let $p_{e|y_T} = \Pr[\mathrm{ext}_p(\mathbf{Z}|\mathbf{Y}_T = y_T) = e]$.

Using the above estimates, we have

$$\begin{aligned}
|\mathrm{ext}_{p,long}(\mathbf{Z}) - U_m| &\leq \sum_{y_T} p_{y_T} |\mathrm{SExt}(\mathbf{X}_S \circ y_T, \mathrm{ext}_p(\mathbf{Z}|\mathbf{Y}_T = y_T)) - U_m| \\
&\leq \left(\sum_{y_T \in Good_y} p_{y_T} |\mathrm{SExt}(\mathbf{X}_S \circ y_T, \mathrm{ext}_p(\mathbf{Z}|\mathbf{Y}_T = y_T)) - U_m| \right) + 2^{-\Omega(n)} \\
&\leq \sum_{y_T \in Good_y} p_{y_T} \left(\sum_e p_{e|y_T} |\mathrm{SExt}(\mathbf{X}_S \circ y_T, e) - U_m| + n^{-\lambda+1} \right) + 2^{-\Omega(n)} \\
&\leq \left(\sum_{y_T \in Good_y} p_{y_T} |\mathrm{SExt}(\mathbf{X}_S \circ y_T, U_d) - U_m| \right) + n^{-\Omega(1)} \\
&= n^{-\Omega(1)}.
\end{aligned}$$

where the last line follows from the fact that $\mathbf{X}_S$ has min-entropy at least $\delta n \log p$. □

10.4.4 One Bit Extractors for 2-Interleaved Sources on $\mathbb{F}_p^{2n}$ with Exponentially Small Error

Note that all our extractor constructions so far have polynomially small error if we insist that the output of the extractor is a bit string. Here we show how to achieve exponentially small error for 2-interleaved sources on $\mathbb{F}_p$, for any large enough prime. However we can output only 1 bit.

Theorem 10.4.7. *There exists $c > 0$ such that for any $\delta > 0$ and any prime $p > 2^{\frac{c}{\delta}}$, there exists an explicit function $\mathrm{ext}_{1bit} : \mathbb{F}_p^{2n} \to \{0,1\}$, such that if $\mathbf{X}$ and $\mathbf{Y}$ are independent sources on $\mathbb{F}_p^n$ with min-entropy k_1, k_2 respectively, satisfying $k_1 > (\frac{1}{2} + \delta) n \log p$ and $k_2 > (5 \log n \log p)/\delta$, $t : [2n] \to [2n]$ is any injective map, then*

$$|\mathrm{ext}_{1bit}((\mathbf{X} \circ \mathbf{Y})_t) \circ \mathbf{X} - U_1 \circ \mathbf{X}| = 2^{-\Omega(k_2)}.$$

Proof. Let $S = \{v_1, \ldots, v_{2n}\}$ be an explicit (n,n-C)-spanning set in $\mathbb{F}_p^{\bar{n}}$ from Lemma 10.3.9. Define

the functions:

$$\mathrm{enc}(z) = \sum_{i=1}^{2n} z_i v_i, \quad \mathrm{ext}(z) = \mathrm{QR}\left(\mathrm{enc}(z)\right),$$

where QR is the quadratic character of $\mathbb{F}^*_{p^{\bar{n}}}$. The proof now follows using Lemma 10.4.1. □

10.4.5 Semi-Explicit Extractors for 2-Interleaved Sources with Linear Output Length and Exponentially Small Error

We note that the extractors constructed so far have either achieved linear output length or exponentially small error, but not both simultaneously. We show that if we allow the extractors to run in sub-exponential time, then we can indeed construct such extractors. (Note that the trivial algorithm to find such an extractor runs in doubly exponential time.) The non-polynomial running time comes from having to compute the discrete logarithm. To reduce the running time, we can in fact use a heuristic algorithm for finding discrete logarithm [BGJT14], which runs in time $n^{O(\log n)}$ on fields of small characteristics under plausible assumptions.

Theorem 10.4.8. *For some $\delta > 0$, there exists a semi-explicit function $\mathrm{ext} : \{0,1\}^{2n} \to \{0,1\}^m$, such that if $\mathbf{X}$ and $\mathbf{Y}$ are independent sources on $\mathbb{F}_2^n$ with min-entropy k_1, k_2 respectively satisfying $k_1 > (1-\delta)n$ and $k_2 > \frac{10}{\delta}\max\{\log n, m\}$, $t : [2n] \to [2n]$ is any permutation, then*

$$|\mathrm{ext}((\mathbf{X} \circ \mathbf{Y})_t) \circ \mathbf{X} - U_m \circ \mathbf{X}| = 2^{-\Omega(k_2)}.$$

Proof. Let $S = \{v_1, \ldots, v_{2n}\}$ be an explicit $(n, n-C)$-spanning set in $\mathbb{F}_2^{\bar{n}}$ constructed using Lemma 10.3.8. Let $m = \frac{\delta k_2}{2}$. For any $z \in \mathbb{F}_p^{2n}$, define the functions:

$$\mathrm{enc}(z) = \sum_{i=1}^{2n} z_i v_i, \quad \mathrm{ext}_1(z) = \log_g(\mathrm{enc}(z)), \quad \mathrm{ext}(z) = \mathrm{ext}_1(z) \pmod{2^m}$$

where g is a generator of $\mathbb{F}^*_{2^{\bar{n}}}$. The proof now follows using Lemma 10.4.1 and Lemma 2.6.2. □

Using the $(n, n-C)$-spanning sets from Lemma 10.3.9 to encode the sources, we obtain the

following theorem using Lemma 10.4.1.

Theorem 10.4.9. *There exists $c > 0$ such that for any $\delta > 0$ and any prime $p > 2^{\frac{c}{\delta}}$, there exists a semi-explicit function* ext $: \mathbb{F}_p^{2n} \to \{0,1\}^m$*, such that if* $\mathbf{X}$*,* $\mathbf{Y}$ *are independent sources on $\mathbb{F}_p^n$ with min-entropy k_1, k_2 respectively satisfying $k_1 > (\frac{1}{2} + \delta)n \log p$ and $k_2 > \frac{5}{\delta} \max\{\log n \log p, m\}$, $t : [2n] \to [2n]$ is any permutation, then*

$$|\text{ext}((\mathbf{X} \circ \mathbf{Y})_t) \circ \mathbf{X} - U_m \circ \mathbf{X}| = 2^{-\Omega(k_2)}.$$

10.4.6 Extractors for 2-Interleaved Sources with Linear Min-Entropy Under the Generalized Paley Graph Conjecture

In this section, we show how to construct extractors for sources with linear min-entropy under the widely believed Generalized Paley Graph Conjecture.

Generalized Paley Graph Conjecture. *Let χ be any non-principal multiplicative character of $\mathbb{F}_{p^n}^*$. For any constant $\delta > 0$, and arbitrary subsets $A, B \subseteq \mathbb{F}_{p^n}$ satisfying $|A|, |B| > p^{\delta n}$, we have*

$$\left| \sum_{a \in A, b \in B} \chi(a + b) \right| \leq p^{-\gamma(\delta)n} |A||B|.$$

Assuming the above conjecture, we obtain the following improved version of Lemma 10.4.1.

Lemma 10.4.10. *Assume the Generalized Paley graph Conjecture. Fix any $\delta > 0$ and any prime p. Let $\mathbf{Z} = (\mathbf{X} \circ \mathbf{Y})_t$ be any 2-interleaved source on $\mathbb{F}_p^{2n}$, where $\mathbf{X}$ and $\mathbf{Y}$ are independent sources on $\mathbb{F}_p^n$ with min-entropy k_1 and k_2 respectively, and $t : [2n] \to [2n]$ is any permutation. Also suppose χ is any nonprincipal multiplicative character of $\mathbb{F}_{p^{\bar{n}}}^*$ and* enc *is an arbitrary (n, s)-encoding from $\mathbb{F}_p^{2n}$ to $\mathbb{F}_p^{\bar{n}}$. Then, there exists $\gamma = \gamma(\delta)$ such that*

$$\mathbf{E}_{\mathbf{X}} |\mathbf{E}_{\mathbf{Y}}[\chi(\text{enc}(\mathbf{Z}))]| \leq p^{-\gamma n},$$

whenever

- $k_1 \geq \delta \bar{n} \log p + (n-s) \log p$, *and*
- $k_2 \geq \delta \bar{n} \log p + (n-s) \log p$.

Proof. For any $z \in \mathbb{F}_p^{2n}$, let

$$\text{enc}(z) = \sum_{i=1}^{2n} z_i v_i$$

where $S = \{v_1, \ldots, v_{2n}\} \subset \mathbb{F}_p^{\bar{n}}$ is (n, s)-spanning.

We have,

$$\chi(\text{enc}(\mathbf{Z})) = \chi\left(\sum_{i=1}^{2n} \mathbf{Z}_i v_i\right) = \chi\left(\sum_{i=1}^{n} \mathbf{X}_i v_{t(i)} + \sum_{j=1}^{n} \mathbf{Y}_j v_{t(n+j)}\right)$$

Define the following independent sources:

$$\mathbf{X}' = \sum_{i=1}^{n} x_i v_{t(i)} : x \sim \mathbf{X}, \qquad \mathbf{Y}' = \sum_{j=1}^{n} y_j v_{t(n+j)} : y \sim \mathbf{Y}.$$

Using Lemma 2.3.12, it follows that: $H_\infty(\mathbf{X}') \geq k_1 - (n-s)\log p$ and $H_\infty(\mathbf{Y}') \geq k_2 - (n-s)\log p$.

Thus, we have

$$\begin{aligned}
\mathbf{E}_{\mathbf{X}} |\mathbf{E}_{\mathbf{Y}}[\chi(\text{enc}(\mathbf{Z}))]| &= \mathbf{E}_{x \sim \mathbf{X}} \left| \mathbf{E}_{y \sim \mathbf{Y}} \left[\chi\left(\sum_{i=1}^{n} x_i v_{t(i)} + \sum_{j=1}^{n} y_j v_{t(n+j)} \right) \right] \right| \\
&= \mathbf{E}_{\mathbf{X}'} \left| \mathbf{E}_{\mathbf{Y}'} \left[\chi\left(\mathbf{X}' + \mathbf{Y}' \right) \right] \right| \\
&\leq p^{-\gamma n}
\end{aligned}$$

where the last inequality follows using the Generalized Paley Graph Conjecture. □

Using the above lemma, we have the following theorem.

Theorem 10.4.11. *Assume the Generalized Paley Graph Conjecture. For any $\delta, \lambda > 0$, there exists an explicit function $\mathrm{ext}_{conjecture} : \{0,1\}^{2n} \to \{0,1\}^m$, $m = \lambda \log n$, such that if $\mathbf{X}$ and $\mathbf{Y}$ are independent sources with min-entropy δn each, and $t : [2n] \to [2n]$ is any permutation, then*

$$|\mathrm{ext}_{conjecture}((\mathbf{X} \circ \mathbf{Y})_t) - U_m| = n^{-\Omega(1)}.$$

Proof. Let $S = \{v_1, \ldots, v_{2n}\}$ be an explicit $(n, n-C)$-spanning set in $\mathbb{F}_p^{\bar{n}}$ constructed using Lemma 10.3.8. Further, as in the proof of Theorem 10.4.2, we choose the rate of the code in Lemma 10.3.9 such that $m|\bar{n}$ and $m = \lambda \log n$. Let $M = n^\lambda$. For any $z \in \mathbb{F}_2^{2n}$, define the functions:

$$\mathrm{enc}(z) = \sum_{i=1}^{2n} z_i v_i, \quad \mathrm{ext}_1(z) = (\mathrm{enc}(z))^{\frac{p^{\bar{n}}-1}{M}}, \quad \mathrm{ext}(z) = \log_g(\mathrm{ext}_1(z))$$

where g is a generator of $G = \{x^{\frac{2^{\bar{n}}-1}{M}} : x \in \mathbb{F}_{2^{\bar{n}}}^*\}$. The proof now follows using Lemma 10.4.10 and Lemma 2.6.1. □

We note that assuming the above conjecture, the output length of the above extractor can be improved to $\Omega(n)$ if both $\mathbf{X}$ and $\mathbf{Y}$ have min-entropy rate more than $\frac{1}{4}$ by using the proof method of Theorem 10.4.6.

10.5 Interleaved Non-Malleable Extractors

In this section, we show that the proof technique developed in constructing extractors for 2-interleaved sources can be used to construct non-malleable extractors in the interleaved model.

Theorem 10.5.1. *There exists $\lambda_1 > 0$ such that for any $\delta, \lambda_2 > 0$, $c > c(\delta)$ and any prime $p > 2^{\frac{\lambda_1}{\delta}}$, there exists an explicit function $\mathrm{nmExt} : \mathbb{F}_p^{2n} \to \{0,1\}^m$, $m = \lambda_2 \log n$, such that if $\mathbf{X}$, $\mathbf{Y}$ are independent sources on $\mathbb{F}_p^n$ with min-entropy k_1, k_2 respectively, satisfying $k_1 > (\frac{1}{2} + \delta) n \log p$ and $k_2 > c \max\{m, \log n\}$, $t : [2n] \to [2n]$ is any injective map and $f : \mathbb{F}_p^n \to \mathbb{F}_p^n$ is any function*

with no fixed points, then

$$|\mathrm{nmExt}((\mathbf{X}\circ\mathbf{Y})_t)\circ\mathrm{nmExt}((\mathbf{X}\circ f(\mathbf{Y}))_t)\circ\mathbf{Y} - U_m\circ\mathrm{nmExt}((\mathbf{X}\circ f(\mathbf{Y}))_t)\circ\mathbf{Y}| = n^{-\Omega(1)}.$$

To prove the above theorem, we recall a character sum estimate of Dodis et al. [DLWZ14].

Theorem 10.5.2. *For any $\delta > 0$ and $\eta < \frac{1}{2}$, suppose S and T are non-empty subsets of $\mathbb{F}_q$ satisfying $|S| > q^{\frac{1}{2}+\delta}$ and $|T| > \max\{(\frac{1}{\eta})^{\frac{7}{\delta}}, (\log q)^8\}$. Let $f : \mathbb{F}_q \to \mathbb{F}_q$ be any arbitrary function with no fixed points. For arbitrary multiplicative characters χ_a and χ_b, such that χ_a is nonprincipal, we have*

$$\sum_{y\in T}\left|\sum_{x\in S}\chi_a(x+y)\chi_b(x+f(y))\right| < \eta|S||T|.$$

Proof of Theorem 10.5.1. We use encoding based on spanning vectors. In particular, let $S = \{v_1,\ldots,v_{2n}\}$ be an explicit $(n, n-C)$-spanning set in $\mathbb{F}_p^{\bar{n}}$ constructed using Lemma 10.3.9. Further, as in the proof of Theorem 10.4.2, we choose the rate of the code in Lemma 10.3.9 such that $m|\bar{n}$ and $m = \lambda_2 \log_p n$. Let $M = n^{\lambda_2}$. For any $z \in \mathbb{F}_p^{2n}$, define the functions:

$$\mathrm{enc}(z) = \sum_{i=1}^{2n} z_i v_i, \quad \mathrm{ext}_1(z) = (\mathrm{enc}(z))^{\frac{p^{\bar{n}}-1}{M}}, \quad \mathrm{ext}(z) = \log_g(\mathrm{ext}_1(z))$$

where g is a generator of $G = \{x^{\frac{p^{\bar{n}}-1}{M}} : x \in \mathbb{F}_{p^{\bar{n}}}^*\}$. We prove the following claim.

Claim 10.5.3. *Let ψ_a and ψ_b be arbitrary characters of the additive group $\mathbf{Z}_M$ such that ψ_a is nontrivial. Then,*

$$\mathbf{E}_{y\sim\mathbf{Y}}\,|\mathbf{E}_{x\sim\mathbf{X}}[\psi_a(\mathrm{nmExt}((\mathbf{X}\circ\mathbf{Y})_t))\psi_b(\mathrm{nmExt}((\mathbf{X}\circ f(\mathbf{Y}))_t))]| = n^{-\Omega(1)}.$$

Before proving this claim, we note that Theorem 10.5.1 follows directly from Claim 10.5.3 by using Lemma 2.6.3.

Proof of Claim 10.5.3. Let $t([n]) = T_1$ and $t([n+1,2n]) = T_2$. Since S is (n,n)-spanning, it follows

that the set $\{v_i : i \in T_1\}$ consists of linearly independent vectors. Similarly $\{v_j : j \in T_2\}$ is a set of linearly independent vectors.

Let $\psi_a(x) = e_M(ax)$, where $a \neq 0 \pmod{M}$. Also let $\psi_b(x) = e_M(bx)$. If $b = 0 \pmod{M}$, the claim follows from Lemma 10.4.1. Thus suppose $b \neq 0 \pmod{M}$.

We have,

$$\begin{aligned}\psi_a(\mathrm{nmExt}((\mathbf{X} \circ \mathbf{Y})_t) &= e_M(a \log_g(\mathrm{ext}_1((\mathbf{X} \circ \mathbf{Y})_t))) \\ &= \chi_a\left(\sum_{i=1}^{n} \mathbf{X}_i v_{t(i)} + \sum_{j=1}^{n} \mathbf{Y}_j v_{t(n+j)}\right) \\ &= \chi_a\left(\mathbf{X}' + \mathbf{Y}'\right)\end{aligned}$$

where $\chi_a(x) = e_M(a \log_g(x))$ is a nonprincipal multiplicative character of $\mathbb{F}^*_{p^{\bar{n}}}$ of order $\frac{M}{\gcd(M,a)}$, $\mathbf{X}' = \sum_{i=1}^{n} x_i v_{t(i)} : x \sim \mathbf{X}$ and $\mathbf{Y}' = L(\mathbf{Y})$, $L : \mathbb{F}_p^n \to \mathbb{F}_p^{\bar{n}}$ being the injective linear map:

$$L(y) = \sum_{j=1}^{n} y_j v_{t(n+j)}.$$

Further,

$$\begin{aligned}\psi_b(\mathrm{nmExt}((\mathbf{X} \circ f(\mathbf{Y}))_t) &= e_M(b \log_g(\mathrm{ext}_1((\mathbf{X} \circ \mathbf{Y})_t))) \\ &= \chi_b\left(\sum_{i=1}^{n} \mathbf{X}_i v_{t(i)} + \sum_{j=1}^{n} f(\mathbf{Y})_j \mathbf{Y}_{t(n+j)}\right) \\ &= \chi_b\left(\mathbf{X}' + f'(\mathbf{Y}')\right)\end{aligned}$$

where $f' = L \circ f \circ L^{-1}$ and $\chi_b(x) = e_M(b \log_g(x))$ is a nonprincipal multiplicative character of $\mathbb{F}^*_{p^{\bar{n}}}$ of order $\frac{M}{\gcd(M,b)}$.

We claim that f' has no fixed points. This can be proved in the following way. Suppose $f'(x) = x$ for some x. This implies that $f(L^{-1}(x)) = L^{-1}(x)$ and hence $f(w) = w$ for $w = L^{-1}(x)$.

This contradicts our assumption on f. Thus f' has no fixed points.

It now follows from Theorem 10.5.2 that

$$\mathbf{E}_{x' \sim \mathbf{X}'} |\mathbf{E}_{y' \sim \mathbf{Y}'} [\chi_a(x' + y') \chi_b(x' + f'(y'))]| = n^{-\Omega(1)}.$$

□

□

If we allow the non-malleable extractor to run in sub-exponential time, then using the proof method of the above theorem, it can be shown that the extractor from Theorem 10.4.9 is non-malleable. Thus, we have the following result.

Theorem 10.5.4. *There exists $\lambda > 0$ such that for any $\delta > 0$, $c > c(\delta)$ and any prime $p > 2^{\frac{\lambda}{\delta}}$, there exists a semi-explicit function* $\mathrm{nmExt} : \mathbb{F}_p^{2n} \to \{0,1\}^m$*, $m = \Omega(n)$, such that if $\mathbf{X}$, $\mathbf{Y}$ are independent sources on $\mathbb{F}_p^n$ with min-entropy k_1, k_2 respectively, satisfying $k_1 > (\frac{1}{2} + \delta) n \log p$ and $k_2 > c \max\{m, \log n\}$, $t : [2n] \to [2n]$ is any permutation and $f : \mathbb{F}_p^n \to \mathbb{F}_p^n$ is any function with no fixed points, then*

$$|\mathrm{nmExt}((\mathbf{X} \circ \mathbf{Y})_t) \circ \mathrm{nmExt}((\mathbf{X} \circ f(\mathbf{Y}))_t) \circ \mathbf{Y} - U_m \circ \mathrm{nmExt}((\mathbf{X} \circ f(\mathbf{Y}))_t) \circ \mathbf{Y}| = 2^{-\Omega(k_2)}.$$

We note that under the Generalized Paley Graph Conjecture, we can reduce the min-entropy requirement of the source $\mathbf{X}$ in Theorem 10.5.1 to βn, for any constant $\beta > 0$.

10.6 Proof of Theorem 25

We briefly recall some definitions from communication complexity. We refer the reader to [KN97] for more background. For convenience, we define boolean functions with range $\{-1, 1\}$ (instead of $\{0, 1\}$).

Definition 10.6.1. *Let $f : [p]^{2n} \to \{-1, 1\}$ be any function. Fix any equi-partition of $[2n]$ into subsets S, T. For any rectangle R and probability distribution μ on $[p]^{2n}$, denote*

$$\mathrm{Disc}^{\mu,R}_{S,T}(f) = |\Pr_{\mu}[f(x_S, y_T) = 1 \text{ and } (x, y) \in R] - \Pr_{\mu}[f(x_S, y_T) = -1 \text{ and } (x, y) \in R]|.$$

Definition 10.6.2. *The discrepancy of $f : [p]^{2n} \to \{-1, 1\}$ with respect to an equi-partition of [2n] into S, T and distribution μ on $[p]^{2n}$ is defined as:*

$$\mathrm{Disc}^{\mu}_{S,T}(f) = \left\{ \max_{R} \left(\mathrm{Disc}^{\mu,R}_{S,T}(f) \right) \right\}.$$

Definition 10.6.3. *The maximal-equipartition discrepancy of $f : [p]^{2n} \to \{0, 1\}$ with respect to a distribution μ on $[p]^{2n}$ is defined as:*

$$\mathrm{Disc}^{\mu}_{best}(f) = \max_{\substack{S,T:|S|=|T|=n, \\ S \cup T = [2n]}} \left\{ \mathrm{Disc}^{\mu}_{S,T}(f) \right\}.$$

The following theorem provides a method to lower bound randomized best-paritition communication complexity of f using its maximal-equi-partition discrepancy. A proof can be found in [KN97].

Theorem 10.6.4. *For every function $f : [p]^{2n} \to \{-1, 1\}$, every probability distribution μ on $[p]^{2n}$ and every $\epsilon \geq 0$,*

$$R^{best, \frac{1}{2} - \epsilon}(f) \geq \log \left(\frac{2\epsilon}{\mathrm{Disc}^{\mu}_{best}(f)} \right).$$

We now prove Theorem 25.

Proof of Theorem 25. We show that the explicit extractor from Theorem 10.4.7 is the required function. Recall the construction of the extractor.

Let $S = \{v_1, \ldots, v_{2n}\}$ be an explicit (n,n-C)-spanning set in $\mathbb{F}_p^{\bar{n}}$ constructed using Lemma 10.3.9, $\bar{n} = n(1 + 2\delta)$.

Define the functions:

$$\mathrm{enc}(z) = \sum_{i=1}^{2n} z_i v_i, \qquad \mathrm{ext}(z) = \mathrm{QR}\left(\mathrm{enc}(z)\right),$$

where QR is the quadratic character of $\mathbb{F}^*_{p^{\bar{n}}}$.

We claim that the randomized best partition discrepancy of ext with error $\frac{1}{2} - p^{-\gamma n}$ is at least $(\frac{1}{4} - \delta - \gamma) n \log p$.

Let μ be the uniform distribution on $[p]^{2n}$.

Claim 10.6.5. *For any equi-partition of $[2n]$ into disjoint subsets S and T,*

$$\log\left(\frac{1}{\mathrm{Disc}^{\mu}_{S,T}(\mathrm{ext})}\right) \geq \left(\frac{1}{4} - \delta\right) n \log p.$$

We note that the proof of Theorem 25 is direct from Claim 10.6.5 by using Theorem 10.6.4.

Proof of Claim 10.6.5. Fix any rectangle $R = X \times Y$, for arbitrary subsets $X, Y \subseteq [p]^n$. We use $\mathbf{X}$ to denote the flat distribution supported on the sets X (and similarly ley $\mathbf{Y}$ denote the flat distribution on Y). We have,

$$\mathrm{Disc}^{\mu,R}_{S,T}(ext) = \frac{|X||Y|}{p^{2n}} \left|\mathbf{E}_{x\in\mathbf{X}, y\in\mathbf{Y}}[QR\left(\mathrm{enc}(x_S \circ y_T)\right)]\right|$$

We note that if $|X| \leq p^{\frac{3n}{4}}$ or $|Y| \leq p^{\frac{3n}{4}}$, the claim follows easily.

Thus suppose $|X|, |Y| > p^{\frac{3n}{4}}$. Define the distribution $\mathbf{Z} = (\mathbf{X} \circ \mathbf{Y})_\pi$, where $\pi : [2n] \to [2n]$ is a permutation defined in the following way: Let $S = \{s_1, \ldots, s_n\}$ and $T = \{t_1, \ldots, t_n\}$ such that $s_1 \leq \ldots \leq s_n$ and $t_1 \leq \ldots \leq t_n$. For any $i \in [n]$, define $\pi(i) = s_i$ and for any $j \in [n+1, 2n]$, define $\pi(j) = t_j$ (thus, $\pi([n]) = S$ and $\pi([n+1, 2n]) = T$).

We note that enc is an (n, n)-encoding from $\mathbb{F}_p^{2n} \to \mathbb{F}_p^{\bar{n}}$. Thus,

$$\mathrm{enc}(\mathbf{Z}) = \mathbf{X}' + \mathbf{Y}',$$

where $\mathbf{X}'$ and $\mathbf{Y}'$ are independent sources on $\mathbb{F}_p^{\bar{n}}$ with $H_\infty(\mathbf{X}') = \log(|X|)$ and $H_\infty(\mathbf{Y}') = \log(|Y|)$.

Using Theorem 2.5.4, with $\lambda = 1$, we have

$$\left|\mathbf{E}[QR\left(\mathbf{X}' + \mathbf{Y}'\right))]\right| \leq 2\left(\left(\frac{p^{\bar{n}}}{|\mathbf{X}||\mathbf{Y}|}\right)^{\frac{1}{2}} + \left(\frac{p^{\frac{\bar{n}}{2}}}{|\mathbf{X}|}\right)^{\frac{1}{2}}\right)$$

Thus,

$$\begin{aligned}
\mathrm{Disc}_{S,T}^{\mu,R}(ext) &\leq 2\left(\frac{|X||Y|}{p^{2n}}\right)\left(\left(\frac{p^{\bar{n}}}{|X||Y|}\right)^{\frac{1}{2}} + \left(\frac{p^{\frac{\bar{n}}{2}}}{|X|}\right)^{\frac{1}{2}}\right) \\
&\leq 2\left(\frac{|X|^{\frac{1}{2}}|Y|^{\frac{1}{2}}}{p^{2n-\frac{\bar{n}}{2}}} + \frac{|\mathbf{X}|^{\frac{1}{2}}}{p^{n-\frac{\bar{n}}{4}}}\right) \\
&\leq 2(p^{-(n-\frac{\bar{n}}{2})} + p^{-\frac{n}{2}+\frac{\bar{n}}{4}})
\end{aligned}$$

Since the above estimate holds for any arbitrary rectangle R, we have

$$\log\left(\frac{1}{\mathrm{Disc}_{S,T}^{\mu}(\mathrm{ext})}\right) \geq \left(\frac{1}{4} - \delta\right) n \log p.$$

□

□

p	k_1	k_2	Output Length	Error	Reference	Remarks
2	$\geq (1-\beta)n$	$\geq (1-\beta)n$	γn, $\gamma < \beta$	$2^{-\Omega(n)}$	[RY11]	Not strong
2	$\geq (1-\alpha)n$	$\geq 35\lambda \log n$	$\lambda \log n$	$n^{-\alpha}$	Theorem 23	Strong in $\mathbf{X}$
2	$\geq (1-\alpha)n$	$\geq 35\lambda \log n$	Output in $\mathbf{Z}_M$, $M = n^\lambda$	$2^{-\Omega(k_2)}$	Theorem 24	Strong in $\mathbf{X}$
any $p > 2^{\frac{c}{\delta}}$	$\geq (\frac{1}{2} + \delta)n \log p$	$\geq c_1(\delta, \lambda, p) \log n$	$\lambda \log n$	$n^{-\alpha}$	Theorem 10.4.5	Strong in $\mathbf{X}$
any $p > 2^{\frac{c}{\delta}}$	$\geq (\frac{1}{2} + \delta)n \log p$	$\geq (\frac{1}{2} + \delta)n \log p$	$\Omega(n)$	$n^{-\alpha}$	Theorem 10.4.6	Not strong
any $p > 2^{\frac{c}{\delta}}$	$\geq (\frac{1}{2} + \delta)n \log p$	$\geq c_2(\delta, \lambda, p) \log n$	1 bit	$2^{-\Omega(k_2)}$	Theorem 10.4.7	Strong in $\mathbf{X}$
any $p > 2^{\frac{c}{\delta}}$	$\geq (\frac{1}{2} + \delta)n \log p$	$\geq c_1(\delta, \lambda, p)\lambda \log n$	$\Omega(k_2)$	$2^{-\Omega(k_2)}$	Theorem 10.4.9	Semi-explicit construction
2	$\geq \gamma n$, any constant γ	$\geq \gamma n$	$\lambda \log n$	$n^{-\alpha}$	Theorem 10.4.11	Assuming Generalized Paley Graph Conjecture

Table 10.1: Results on Extractors for 2-Interleaved Sources. The setting is as follows: $\mathbf{Z} = (\mathbf{X} \circ \mathbf{Y})_t$ is an arbitrary 2-interleaved source on $[p]^{2n}$, where $\mathbf{X}$ and $\mathbf{Y}$ are independent sources on $[p]^n$ (for some prime p) with min-entropy k_1 and k_2 respectively, and $t : [2n] \to [2n]$ is an arbitrary permutation. Let α be a small enough constant and c a large enough constant. Also let $\lambda > 1$ be any constant. We also list the result of [RY11] in Table 10.1.

Chapter 11

Seedless Non-Malleable Extractors

[1] Cheraghchi and Guruswami [CG14b] introduced seedless non-malleable extractors as a natural generalization of seeded non-malleable extractors (see Chapter 4 for more details on seeded non-malleable extractors). They also showed a way to construct non-malleable codes from efficient constructions of such seedless non-malleable extractors. Informally, non-malleable codes are a generalization of error-detecting codes to handle a much larger class of tampering functions (rather than just bit erasure or modification). We refer the reader to Chapter 12 for more details and our results on non-malleable codes. Thus apart from a natural notion, the applications to non-malleable codes provides further motivation for explicitly constructing seedless non-malleable extractors.

Definition 11.0.1 (Seedless C-Non-Malleable Extractor)**.** *A function* $\mathrm{nmExt} : (\{0,1\}^n)^C \to \{0,1\}^m$ *is a seedless* C*-non-malleable extractor for min-entropy* k *and error* ϵ *if it satisfies the following property: Let* $\mathbf{X}_1, \ldots, \mathbf{X}_C$ *be independent* (n,k)*-sources and for each* $i \in [C]$*, let* $f_i : \{0,1\}^n \to \{0,1\}^n$ *be arbitrary tampering functions, such that at least one* f_i *has no fixed points. Then,*

$$|\mathrm{nmExt}(\mathbf{X}_1, \ldots, \mathbf{X}_C), \mathrm{nmExt}(f_1(\mathbf{X}_1), \ldots, f_C(\mathbf{X}_C)) - \mathbf{U}_m, \mathrm{nmExt}(f_1(\mathbf{X}_1), \ldots, f_C(\mathbf{X}_C))| < \epsilon.$$

[1] parts of this chapter have been previously published [CZ14, CGL16]

In [CG14b], it was left as an open problem to construct a seedless C-non-malleable extractor even for $k = n$, for any $C = o(n)$.

Using the probabilistic method, Cheraghchi and Guruwsami [CG14b] showed the existence of seedless 2-non-malleable extractor for $k = \Omega(\log n)$ and $\epsilon = 2^{-\Omega(k)}$ with $m = \Omega(k)$. However giving explicit constructions turns out to be tricky, even for $k = n$, and there were no known constructions prior to work in this thesis. More specifically, it appears nontrivial to extend existing constructions of seeded non-malleable extractors when both sources are tampered. For example, for sources on $\mathbb{F}_p$, the 2-source extractor from Lemma 2.5.4: $\chi(x + y)$, where χ is the quadratic character was shown to be a seeded non-malleable extractor [DLWZ14]. However it fails to work against tampering functions $f(x) = x + 1$ and $g(y) = y - 1$, even for full entropy.

11.1 Our Results

The results in this chapter are based on joint works with Vipul Goyal, Xin Li, and David Zuckerman [CZ14, CGL16]. We provide two different constructions of seedless non-malleable extractors, which rely on very different set of techniques.

Our first construction however requires access to 10 sources but has the advantage the the output length is $\Omega(k)$ and error $2^{-\Omega(n)}$. In fact, as we will see in Chapter 12, we use this extractor to give the first explicit constructions of non-malleable codes with constant rate. This construction relies on techniques from the area of additive combinatorics.

Theorem 27. *For some $\delta > 0$ there exists a polynomial time construction of a (k, ϵ)-seedless non-malleable extractor for 10 independent sources* $\text{nmExt} : (\{0,1\}^n)^{10} \to \{0,1\}^m$ *with $k = (1 - \delta)n$, $\epsilon = 2^{-\Omega(n)}$ and $m = \Omega(k)$.*

We present the proof of Theorem 27 in Section 11.2.

Our second construction resolves the open problem posed by Cheraghchi and Guruswami, and gives explicit non-malleable extractors for 2 sources. However, the output length of this extractor is polynomially small. An advantage of this construction is that it generalizes to handle

multiple tamperings, which yields generalized non-malleable codes. This construction is based on techniques developed in Chapter 3, and in particular the construction is very similar to the seeded non-malleable extractor construction in Theorem 1. To present our result in full generality, we introduce seedless (C,t)-non-malleable extractors.

Definition 11.1.1 (Seedless (C,t)-Non-Malleable Extractor)**.** *A function* $\text{nmExt} : (\{0,1\}^n)^C \to \{0,1\}^m$ *is a seedless* C*-non-malleable extractor for min-entropy* k *and error* ϵ *if it satisfies the following property: Let* $\mathbf{X}_1, \ldots, \mathbf{X}_C$ *be independent* (n,k)*-sources. Further, for each* $i \in [C], j \in [t]$*, let* $f_{i,j} : \{0,1\}^n \to \{0,1\}^n$ *be an arbitrary tampering function, such that for each* $j \in [t]$*, at least one* $f_{i,j}$ *has no fixed points. Then,*

$$|\text{nmExt}(\mathbf{X}_1, \ldots, \mathbf{X}_C), \text{nmExt}(f_{1,1}(\mathbf{X}_1), \ldots, f_{1,C}(\mathbf{X}_C)), \ldots, \text{nmExt}(f_{t,1}(\mathbf{X}_1), \ldots, f_{t,C}(\mathbf{X}_C))$$
$$-\mathbf{U}_m, \text{nmExt}(f_{1,1}(\mathbf{X}_1), \ldots, f_{1,C}(\mathbf{X}_C)), \ldots, \text{nmExt}(f_{t,1}(\mathbf{X}_1), \ldots, f_{t,C}(\mathbf{X}_C))| < \epsilon.$$

The following is the second main result of this chapter.

Theorem 28. *There exists a constant* $\gamma > 0$ *such that for all* $n > 0$ *and* $t \leq n^\gamma$*, there exists an efficient seedless* $(2,t)$*-NM extractor at min-entropy* $n - n^\gamma$ *with error* $2^{-n^{\Omega(1)}}$ *and output length* $m = n^{\Omega(1)}$*.*

We present the construction in Theorem 28 in Section 11.3.

11.2 An Explicit Seedless Non-Malleable Extractor for 10 Sources

We prove Theorem 27 in this section. We first introduce some tools which are used in our construction.

Notation For a vector $v \in \mathbb{F}_p^n$, we use $\Pi_S(v)$ to denote the projection of v to the coordinates indexed by the elements in $S \subset [n]$. We extend the action of Π_S to sets in the obvious manner. We use Π_i for $\Pi_{\{i\}}$.

11.2.1 Some Results from Additive Combinatorics

We recall some well known results from additive combinatorics. We refer the reader to the excellent book by Tao and Vu [TV06] for more details.

Definition 11.2.1. *For vectors $v, w \in \mathbb{F}_p^n$, where $v = (v_1, \ldots, v_n)$ and $w = (w_1, \ldots, w_n)$, we define*

$$v \odot w = (v_1 w_1, \ldots, v_n w_n)$$

Definition 11.2.2. *For subsets $A, B \subseteq \mathbb{F}_p^n$, define the sets :*

$$A + B = \{a + b : a \in A, b \in B\}$$

$$A \odot B = \{a \odot b : a \in A, b \in B\}$$

Observation 11.2.3. *$(\mathbb{F}_p^*)^n$ is a group under the operation $\odot$.*

Lemma 11.2.4 (Plünnecke-Ruzsa)**.** *Let A, B be finite subsets in an additive group G. Then*

$$|A + A| \leq \frac{|A + B|^4}{|A||B|^2}$$

Lemma 11.2.5 (Plünnecke-Ruzsa)**.** *Let A be a finite subset of any additive group G. Then*

$$|A - A| \leq \left(\frac{|A + A|}{|A|}\right)^3 |A|$$

Lemma 11.2.6 (Balog-Szemerédi-Gowers lemma [BS94, Gow98])**.** *Let A, B be finite subsets of an additive group G and let $|A|^{1-\rho_1} \leq |B| \leq |A|^{1+\rho_1}$. If $\mathrm{cp}(A + B) \geq |A|^{-(1+\rho_2-\rho_1)}$, then there exists subsets $A' \subseteq A$, $B' \subseteq B$ such that $|A'| \geq |A|^{1-10\rho_2}$, $|B'| \geq |B|^{1-10\rho_2}$, and $|A' + B'| \leq |A|^{1+\rho_1+10\rho_2}$.*

11.2.2 Some Known Extractor Constructions

We recall some known results on multi-source extractors and non-malleable extractors. We recall a 3-source extractor constructed in [Rao09a].

Theorem 11.2.7 ([Rao09a]). *For every n and and constant $\delta > 0$ there exists an explicit function* $3\text{ext} : \{0,1\}^n \to \{0,1\}^m$, *$m = \Omega(n)$, such that if $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ are independent $(n, \delta n)$ sources then*

$$|3\text{Ext}(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) - U_m| < 2^{-\Omega(n)}$$

Explicit constructions of seeded non-malleable extractors follow from works of [DLWZ14] and [Li12b]. The output length in [DLWZ14] relies on an unproven but widely believed conjecture on primes while the output length in [Li12b] is unconditional. Further, either of the non-malleable extractors from [DLWZ14] or [Li12b] is also a strong 2-source extractor.

Theorem 11.2.8 ([DLWZ14, Li12b]). *Let $\delta > 0$ be a constant. For all n, there exists an explicit function* $\text{snmExt} : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$, *$m = \Omega(n)$, satisfying: Suppose X, Y are independent sources on $\{0,1\}^n$ with min-entropy k_1, k_2 respectively.*

1. *If $(k_1 + k_2) \geq (1 + \delta)n$, then*

$$|\text{snmExt}(\mathbf{X}, \mathbf{Y}), \mathbf{X} - \mathbf{U}_m, \mathbf{X}| < 2^{-\Omega(n)}, \;\; |\text{snmExt}(\mathbf{X}, \mathbf{Y}), \mathbf{Y} - \mathbf{U}_m, \mathbf{Y}| < 2^{-\Omega(n)}$$

2. *If $k_1, k_2 > (1 - \delta)n$ and f is any tampering function with no fixed points, then*

$$|\text{snmExt}(\mathbf{X}, \mathbf{Y}), \text{snmExt}(\mathbf{X}, f(\mathbf{Y})) - \mathbf{U}_m, \text{snmExt}(\mathbf{X}, f(\mathbf{Y}))| < 2^{-\Omega(n)}$$

11.2.3 A Sum-Product Estimate

We recall a sum-product theorem over prime fields follows from [BKT04, BGK06, Kon03].

Theorem 11.2.9 (Sum-product over prime fields)**.** *Let $\mathbb{F}_p$ be any prime field and let $A \subset \mathbb{F}_p$ be any non-empty subset such that $|A| < p^{1-\delta}$ for some constant $\delta > 0$. Then there exists a constant $\tau = \tau(\delta) > 0$, such that*

$$|A + A| + |A \cdot A| \geq |A|^{1+\tau}$$

An analogue of Theorem 11.2.9 over $\mathbb{F}_p \times \mathbb{F}_p$ was proved by Bourgain in [Bou05a]. We extend this to sets over $\mathbb{F}_p^4$ in the following theorem and use it in our proof of Theorem 3. It is stated in a convenient way.

Theorem 11.2.10. *There exists $\tau_0 > \tau_1 > 0$ such that the following holds: Let A be a subset of $\mathbb{F}_p^4$ satisfying $|A \cap (\mathbb{F}_p^*)^4| \geq \frac{|A|}{2}$. Suppose that for any subset $A_1 \subseteq A$ satisfying $|A_1| \geq p^{-\tau_1}|A|$, the following conditions holds.*

1. *$\Pi_{\{1,2\}}(A_1) \geq p^{1+\tau_0}$ and $\Pi_{\{3,4\}}(A_1) \geq p^{1+\tau_0}$.*
2. *$A_1 \nsubseteq P$, where P is a 2-dimensional linear subspace of $\mathbb{F}_p^4$ of the form*
   (a) *$\{(x_1, x_2, c_1x_1, c_2x_2) : x_1 \in \mathbb{F}_p, x_2 \in \mathbb{F}_p\}$ or*
   (b) *$\{(x_1, x_2, c_2x_2, c_1x_1) : x_1 \in \mathbb{F}_p, x_2 \in \mathbb{F}_p\}$.*

Then there exists a constant $\tau > 0$ (depending on τ_0, τ_1) such that if $|A| < p^{7/3-\tau_1}$, then

$$|A + A| + |A \odot A| > p^{\tau}|A|$$

We present the proof of Theorem 11.2.10 in Section 11.2.7. The proof of Theorem 11.2.10 closely follows and extends the arguments in the sum-product estimate over $\mathbb{F}_p^2$ proved by Bourgain.

Definition 11.2.11. *We call a set A satisfying the conclusion of Theorem* 11.2.10 *to be sum-product friendly. We call a flat distribution sum-product friendly if its support is sum-product friendly.*

11.2.4 A Sum-Product Friendly Encoding

Let τ, τ_0, τ_1 be the constants from Theorem 11.2.10. Let p be any prime satisfying : $p^{\tau_0} > 16$.

Define $\text{enc} : \mathbb{F}_p \to \mathbb{F}_p^2$ in the following way.

$$\text{enc}(x) = (x, x^4 + x^2 + x)$$

Lemma 11.2.12. *Let $S_1, S_2 \subset \mathbb{F}_p$ be subsets of size $p^{1-\delta}$, $p > 3$. Define the distribution*

$$\mathbf{X}_{f,1,2} = (\text{enc}(x_1) + \text{enc}(x_2), \text{enc}(f_1(x_1)) + \text{enc}(f_2(x_2))) : x_1 \sim S_1, x_2 \sim S_2$$

where f_1, f_2 are arbitrary functions.

Then $\mathbf{X}_{f,1,2}$ is $O(p^{-\delta})$-close to a convex combination of at most 4 flat distributions supported on sets of the form

$$T_i = \{(\text{enc}(x_1) + \text{enc}(x_2), \text{enc}(f_1(x_1) + \text{enc}(f_2(x_2))) : (x_1, x_2) \in G_i\},$$

where $G_i \subset \mathbb{F}_p^2$ and $|G_i| = |T_i| \geq p^{2-3\delta}$.

Proof. Let $T \subset \mathbb{F}_p^4$ denote the support of $\mathbf{X}_{f,1,2}$. We partition T into at most 4 parts in the following way.

For any $t \in T$, let $s(t) \subset \mathbb{F}_p^2$ be the set of all $(x_1, x_2) \in S_1 \times S_2$ such that $(\text{enc}(x_1) + \text{enc}(x_2), \text{enc}(f_1(x_1)) + \text{enc}(f_2(x_2))) = t$. Let $r(x)$ denote the cardinality of the set $s(x)$.

We claim that for any $t \in T$, $1 \leq r(x) \leq 4$. The upper bound follows from the following calculation. Let $t = (t_1, t_2, t_3, t_4) \in F_p^4$. Thus for any $(x_1, x_2) \in s(t)$, we have

$$\begin{aligned} x_1 + x_2 &= t_1 \\ x_1^4 + x_1^2 + x_1 + x_2^4 + x_2^2 + x_2 &= t_2 \end{aligned}$$

Substituting for x_2, we have

$$x_1^4 + (t_2 - x_1)^4 + q(x_1, t_1, t_2) = 0$$

where $q(x_1, t_1, t_2)$ has degree at most 2 in x_1. Thus x_1 must satisfy a polynomial of degree exactly 4. For each fixing of x_1, notice that x_2 also gets fixed. Thus $r(t) \leq 4$ for all $t \in T$.

For $i \in [4]$, we define the sets

$$T_i = \{t \in T : r(t) = i\}$$

Thus the T_i's form a partition of T.

Define sets $G_i \subset \mathbb{F}_p^2$, $i \in [4]$, such that for all $t \in T_i$, $|G_i \cap s(t)| = 1$. In other words G_i is constructed by picking exactly one element from $s(t)$ for each $t \in T_i$. Thus $|G_i| = |T_i|$.

We note that for any $t \in T_i$, $\Pr[\mathbf{X}_{f,1,2} = t] = \frac{i}{|S_1||S_2|}$ and hence

$$\Pr[\mathbf{X}_{f,1,2} \in T_i] = \frac{i|G_i|}{|S_1||S_2|}$$

Thus we have

$$\mathbf{X}_{f,1,2} = \sum_i^4 w_i \cdot \big((\mathrm{enc}(x_1) + \mathrm{enc}(x_2), \mathrm{enc}(f_1(x_1)) + \mathrm{enc}(f_2(x_2))) : (x_1, x_2) \sim G_i\big)$$

where $w_i = \frac{i|G_i|}{|S_1||S_2|}$.

For some i, if $|G_i| < p^{2-3\delta}$ then $w_i \leq i \cdot p^{-\delta}$. Thus $\mathbf{X}_{f,i,j}$ is $9 \cdot p^{-\delta}$-close to a distribution $\mathbf{X}'_{f,i,j}$ defined as

$$\mathbf{X}'_{f,1,2} = \sum_i^4 w'_i \cdot \big((\mathrm{enc}(x_1) + \mathrm{enc}(x_2), \mathrm{enc}(f_1(x_1)) + \mathrm{enc}(f_2(x_2))) : (x_1, x_2) \sim G_i\big)$$

where we set w'_i's as follows. Set $w'_i = 0$ for all i such that $w_i < i \cdot p^{-\delta}$. Pick a j such that $w_j \geq j \cdot p^{-\delta}$ and set $w'_j = w_j + \sum_{i:w_i<i\cdot p^{-\delta}} w_i$. For the remaining unset w'_i's, set it equal to w_i. $\square$

Lemma 11.2.13. *Choose a small $\delta_1 > \tau_0$. Let f_1, f_2 be functions with maximum pre-image size bounded by p^{δ_1}. Further assume f_1 has no fixed points. Define the set $A = \{\mathrm{enc}(x_1) +$*

$\mathrm{enc}(x_2), \mathrm{enc}(f_1(x_1)) + \mathrm{enc}(f_2(x_2)) : (x_1, x_2) \in G\}$ *where* $G \subset \mathbb{F}_p^2$ *is a subset of size at least* $p^{1+10\delta_1}$. *Then the set* $A \subset \mathbb{F}_p^4$ *is sum-product friendly.*

Proof. We begin by noting that $p^{1+9\delta_1} < |A| << p^{7/3}$.

We need the following claim.

Claim 11.2.14. *Define the set* $B = \{(\mathrm{enc}(y_1) + \mathrm{enc}(y_2), \mathrm{enc}(g_1(y_1)) + \mathrm{enc}(g_2(y_2))) : (y_1, y_2) \in H\}$ *where* $H \subset \mathbb{F}_p^2$ *is a subset of size at least* $p^{1+10\delta_1}$ *and* g_1, g_2 *are tampering functions with pre-image size bounded by* p^{δ_1}. *Then following inequalities hold :*

- $|B \cap (\{0\} \times \mathbb{F}_p^3)| \leq p$
- $|B \cap (\mathbb{F}_p \times \{0\} \times \mathbb{F}_p^2)| \leq 4.p$
- $|B \cap (\mathbb{F}_p^2 \times \{0\} \times \mathbb{F}_p)| \leq p^{1+\delta_1}$
- $|B \cap (\mathbb{F}_p^3 \times \{0\})| \leq 4p^{1+\delta_1}$

Proof. We have,

$B = \{(y_1+y_2, y_1^4+y_1^2+y_1+y_2^4+y_2^2+y_2, g_1(y_1)+g_2(y_2), g_1(y_1)^4+g_1(y_1)^2+g_1(y_1)+g_2(y_2)^4+ g_2(y_2)^2 + g_2(y_2)) : (y_1, y_2) \in H\}$.

We prove the inequality:

$$|B \cap (\mathbb{F}_p^3 \times \{0\})| \leq 4p^{1+\delta_1}$$

The other inequalities follow using similar arguments.

Fix y_1 to some value in $\mathbb{F}_p$. We note that $g_2(y_2)$ is the root of a monic degree 4 polynomial and hence has at most 4 choices. Thus y_2 can take at most $4p^{\delta_1}$ values by using the bound on the pre-image size of g_2. The inequality now follows by observing that y_1 can take at most p values. □

Using Claim 11.2.14, we have $|A \cap (\mathbb{F}_p^*)^4| \geq (1 - p^{-7\delta_1})|A| > \frac{1}{2}|A|$.

Consider any subset $A_1 \subseteq A$ such that $|A_1| \geq p^{-\tau_1}|A|$. It follows that there exists $G_1 \subseteq G$ such that

$$A_1 = \{(\text{enc}(x_1) + \text{enc}(x_2), \text{enc}(f_1(x_1)) + \text{enc}(f_2(x_2))) : (x_1, x_2) \in G_1\}$$

Thus $|A_1| > p^{1+8\delta_1}$. We also note that $|G_1| \geq |A_1| > p^{1+8\delta_1}$.

We note that $|\Pi_{1,2}(A_1)| = |A_1| > p^{1+\tau_0}$. Further $|\Pi_{3,4}(A_1)| > |A_1|p^{-2\delta_1} > p^{1+6\delta_1} > p^{1+\tau_0}$.

The final part of the proof is to bound the intersection of A_1 with any 2-dimensional linear space P of the forms specified in Theorem 11.2.10.

Suppose $A_1 \subset P = \{(y_1, y_2, c_1y_1, c_2y_2) : y_1, y_2 \in \mathbb{F}_p\}$. Thus we have for all $(x_1, x_2) \in G_1$:

$$f_1(x_1) + f_2(x_2) = c_1(x_1 + x_2)$$
$$f_1(x_1)^4 + f_1(x_1)^2 + f_1(x_1) + f_2(x_2)^4 + f_2(x_2)^2 + f_2(x_2) = c_2(x_1^4 + x_1^2 + x_1 + x_2^4 + x_2^2 + x_2)$$

Fix $x_2 = \alpha$ such that $(x_1, \alpha) \in G_1$ for all $x_1 \in S_1 \subset \mathbb{F}_p$, $|S_1| \geq \frac{|G_1|}{p} \geq p^{8\delta_1}$. Let $f_2(\alpha) = \beta$. We thus have for all $x_1 \in S_1$,

$$f_1(x_1) = c_1x_1 + c_1\alpha - \beta \tag{11.1}$$
$$f_1(x_1)^4 + f_1(x_1)^2 + f_1(x_1) + \beta^4 + \beta^2 + \beta = c_2(x_1^4 + x_1^2 + x_1 + \alpha^4 + \alpha^2 + \alpha) \tag{11.2}$$
$$\tag{11.3}$$

Thus for all $x \in S_1$, the following holds:

$$(c_1x_1 + c_1\alpha - \beta)^4 + (c_1x_1 + c_1\alpha - \beta)^2 + (c_1x_1 + c_1\alpha - \beta) + \beta^4 + \beta^2 + \beta$$
$$-c_2(x_1^4 + x_1^2 + x_1 + \alpha^4 + \alpha^2 + \alpha) = 0 \tag{11.4}$$

To derive a contradiction, we split it into the following cases.

- $c_1 \neq 0$, $c_1\alpha - \beta \neq 0$

In this case notice that the LHS of (11.4) is of degree at least 3 and at most 4 in x_1 and hence can have at most 4 roots, which is a contradiction since $|S_1| \geq p^{8\delta_1} > 4$.

- $c_1 = 0$

  In this case we see that from (11.1), f_1 is constant on S_1 which contradicts the assumption that f_1 has pre-image size at most p^{δ_1}.

- $c_1\alpha - \beta = 0$, $c_1 \neq 0$

  Thus (11.4) simplifies to

$$c_1^4 x_1^4 + c_1^2 x_1^2 + c_1 x_1 + \beta^4 + \beta^2 + \beta - c_2(x_1^4 + x_1^2 + x_1 + \alpha^4 + \alpha^2 + \alpha) = 0 \tag{11.5}$$

  We see that this is at least a linear equation and at most a degree 4 equation in x_1 (and thus a contradiction, as argued above) unless $c_1^4 = c_1^2 = c_1 = c_2$. Thus $c_1 = 1$ (since $c_1 \neq 0$). But by (11.1), we then have $f_1(x_1) = x_1$ for all $x_1 \in S_1$. This contradicts the fact that f_1 has no fixed points.

This contradicts our assumption that $A_1 \subseteq \{(y_1, y_2, c_1y_1, c_2y_2) : y_1, y_2 \in \mathbb{F}_p\}$.

Now suppose $A_1 \subseteq P = \{(y_1, y_2, c_2y_2, c_1y_1) : y_1, y_2 \in \mathbb{F}_p\}$. We arrive at a contradiction using similar arguments as above. We have for all $(x_1, x_2) \in G_1$

$$f_1(x_1) + f_2(x_2) = c_2(x_1^4 + x_1^2 + x_1 + x_2^4 + x_2^2 + x_2)$$
$$f_1(x_1)^4 + f_1(x_1)^2 + f_1(x_1) + f_2(x_2)^4 + f_2(x_2)^2 + f_2(x_2) = c_1(x_1 + x_2)$$

Fix $x_2 = \alpha$ such that $(x_1, \alpha) \in G_1$ for all $x_1 \in S_1 \subset \mathbb{F}_p$, $|S_1| \geq \frac{|G_1|}{|p|} \geq p^{8\delta_1}$. Let $f_2(\alpha) = \beta$. We thus

have for all $x \in S_1$,

$$f_1(x_1) = c_2(x_1^4 + x_1^2 + x_1 + \alpha^4 + \alpha^2 + \alpha) - \beta \tag{11.6}$$

$$f_1(x_1)^4 + f_1(x_1)^2 + f_1(x_1) + \beta^4 + \beta^2 + \beta = c_1(x_1 + x_2) \tag{11.7}$$

$$\tag{11.8}$$

It follows that for all $x \in S_1$,

$$\begin{aligned}\big(c_2(x_1^4 + x_1^2 + x_1 + \alpha^4 + \alpha^2 + \alpha) - \beta\big)^4 + \big(c_2(x_1^4 + x_1^2 + x_1 + \alpha^4 + \alpha^2 + \alpha) - \beta\big)^2 + \\ \big(c_2(x_1^4 + x_1^2 + x_1 + \alpha^4 + \alpha^2 + \alpha) - \beta\big) + \beta^4 + \beta^2 + \beta - c_1(x_1 + \alpha) = 0\end{aligned} \tag{11.9}$$

We note that (11.9) is a degree 16 equation in x_1 (and hence a contradiction since $p^{8\delta_1} > 16$) unless $c_2 = 0$. But if $c_2 = 0$ then from (11.6) we have f_1 is constant on S_1 which contradicts our assumption that f_1 has pre-image size at most p^{δ_1}. This completes our proof that A is sum-product friendly.

□

In the following lemmas, we shall abuse notation and for any set A, we will also use A to denote the flat distribution with support A.

Choose δ_1 small enough such that for a sum-product friendly set A of size $p^{2-5\cdot 10^3\delta_1}$ we have $|A + A| + |A \odot A| > |A|p^{5\cdot 10^4\delta_1}$. This can be ensured by choosing $\delta_1 = 10^{-5} \cdot \tau$, where τ is the constant from Theorem 11.2.10.

Lemma 11.2.15. *Let $G_1, G_2, G_3 \subset \mathbb{F}_p^2$ be subsets of size at least $p^{2-\delta_1}$. Let $f_1, \ldots, f_6$ be functions with pre-image size at most $p^{10\delta_1}$. Further assume f_1 has no fixed points. For $i \in [3]$ define the sets $A_i = \{(\mathrm{enc}(x_{2i-1}) + \mathrm{enc}(x_{2i}), \mathrm{enc}(f_{2i-1}(x_{2i-1})) + \mathrm{enc}(f_{2i}(x_{2i}))) : (x_{2i-1}, x_{2i}) \in G_i\}$. Then $A_1 \odot A_2 + A_3$ is $O(p^{-\delta_1})$-close to a distribution with min-entropy $(2 + 10\delta_1)\log p$.*

To prove the above lemma, we borrow ideas from [BIW06] and use the proof technique

developed in their work.

We begin by proving the following lemmas.

Lemma 11.2.16. *Let $A \subset (\mathbb{F}_p{}^*)^4$, $p^{2-300\delta_1} \leq |A| < p^2$ be such that any subset $A' \subseteq A$ of size greater than $p^{2-5\cdot 10^3\delta_1}$ is sum-product friendly. Supose that for some $B \subset (\mathbb{F}_p^*)^4$, we have $|A \odot B| \leq p^{2+300\delta_1}$, $p^{2-300\delta_1} \leq |B| < p^2$. Then for any $C \subset (\mathbb{F}_p^*)^4$ such that $p^{2-\delta_1} \leq |C| < p^2$, we have* $\mathrm{cp}(A+C) \leq p^{-(2+12\delta_1)}$.

Proof. Since $|A \odot B| \leq p^{2+300\delta_1}$, using Lemma 11.2.4 we have $|A \odot A| \leq |A|p^{2400\delta_1}$. Suppose there is some set C such that $|C| > p^{2-\delta_1}$ and $\mathrm{cp}(A+C) > p^{-(2+12\delta_1)}$. Using Lemma 11.2.6 with $\rho_1 = 200\delta_1$ and $\rho_2 = 220\delta_1$, it follows that there exists sets $A' \subseteq A$, $C' \subseteq C$, $|A'+C'| \leq p^{2+5\cdot 10^3\delta_1}$ and $|A'|, |C'| > p^{2-5\cdot 10^3\delta_1}$. Using Lemma 11.2.4, we get that $|A'+A'| \leq |A'|p^{4\cdot 10^4\delta_1}$. We also have $|A' \odot A'| \leq |A \odot A| \leq |A'|p^{10^4\delta_1}$. By our choice of δ_1, this contradicts A' being sum-product friendly. □

Switching the roles of addition and multiplication gives the following.

Lemma 11.2.17. *Let $A \subset (\mathbb{F}_p{}^*)^4$, $p^{2-300\delta_1} \leq |A| < p^2$ be such that any subset $A' \subseteq A$ of size at least $p^{2-5\cdot 10^3\delta_1}$ is sum-product friendly. Let $B \subset (\mathbb{F}_p^*)^4$ be a set such that $|A+B| \leq p^{2+300\delta_1}$, $p^{2-300\delta_1} \leq |B| < p^2$. Then for any $C \subset (\mathbb{F}_p^*)^4$ such that $p^{2-\delta_1} \leq |C| < p^2$, we have* $\mathrm{cp}(A \odot C) \leq p^{-(2+12\delta_1)}$.

We say that a set is plus-friendly if it satisfies the conclusion of Lemma 11.2.16. Similarly we say that a set is times-friendly if it satisfies the conclusion of Lemma 11.2.17.

Lemma 11.2.18. *Let $A_1 \subset \mathbb{F}_p^4$ be the set defined in Lemma 11.2.15. Then $A_1 = A_+ \cup A_\times \cup A_{11}$ such that the following hold:*

1. *A_+ is empty or plus-friendly*
2. *$A_\times$ is empty or times-friendly*

3. $|A_{11}| \leq |A_1|p^{-\delta_1}$

Proof of Lemma 11.2.18. We start out by replacing A_1 by $A_1 \cap (\mathbb{F}_p^*)^4$. We can do this without loss of generality since as observed in the proof of Lemma 11.2.13, $|A_1 \cap (\mathbb{F}_p^*)^4| > (1 - p^{-\delta_1})|A_1|$ and hence we add the set $A_1 \setminus (\mathbb{F}_p^*)^4$ to A_{11}.

Note that by Lemma 11.2.13, any subset of A_1 of size at least $p^{2-5\cdot 10^3\delta_1}$ is sum-product friendly. Let $A_\times = A_1$ and $A_+ = \emptyset$. We maintain the invariance that A_+ is either plus-friendly or empty. If $A_\times$ is times-friendly then we are done. Else there exists some B of size at least $p^{2-\delta_1}$ such that $\mathrm{cp}(A_\times \odot B) > p^{-(2+12\delta_1)}$. Using Lemma 11.2.6 with $\rho_1 = 2\delta_1$ and $\rho_2 = 14\delta_1$, we have that there exists sets $A' \subseteq A_\times$, $B' \subseteq B$, $|A' \odot B'| \leq p^{2+284\delta_1}$ and $|A'|, |B'| \geq p^{2-282\delta_1}$. Thus, by Lemma 11.2.16, A' is plus-friendly. We remove A' from $A_\times$ and add it to A_+. Further it can be proved that unions of disjoint plus-friendly sets are also plus-friendly. We iterate as above till $A_\times$ is times-friendly or $|A_\times| \leq |A_1|p^{-\delta_1}$. □

Proof of Lemma 11.2.15. By Lemma 11.2.18 we have $A_1 = A_+ \cup A_\times \cup A'$. Using Claim 11.2.14, we have $|A_2 \cap (\mathbb{F}_p^*)^4| > (1 - p^{-\delta_1})|A_2|$ and $|A_3 \cap (\mathbb{F}_p^*)^4| > (1 - p^{-\delta_1})|A_3|$. Thus $A_1 \odot A_2 + A_3$ is $O(p^{-\delta_1})$-close to a convex combination of distributions of the form:

1. $A_+ \odot a_2 + A_3$, $a_2 \in A_2 \cap (\mathbb{F}_p^*)^4$
2. $A_\times \odot A_2 + a_3$, $a_3 \in A_3 \cap (\mathbb{F}_p^*)^4$

By Lemma 11.2.16 and Lemma 11.2.17, we thus have that $A_1 \odot A_2 + A_3$ is $O(p^{-\delta_1})$-close to a distribution with collision probability at most $p^{-(2+12\delta_1)}$. Thus by using Lemma 2.3.3, we have that $A_1 \odot A_2 + A_3$ is $O(p^{-\delta_1})$-close to a distribution with min-entropy $(2 + 10\delta_1)\log p$. □

Theorem 11.2.19. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_8$ be independent sources on $\mathbb{F}_p$ with min-entropy $(1 - \delta)\log p$. Let $f_1, f_2, \ldots, f_8$ be arbitrary functions such that at least one of the f_i's has no fixed points. Further suppose that the pre-image of each f_i is bounded by $p^{10\delta}$. Define the source*

$$\mathbf{X}_{f,i,j} = \mathrm{enc}(\mathbf{X}_i) + \mathrm{enc}(\mathbf{X}_j), \mathrm{enc}(f_i(\mathbf{X}_i)) + \mathrm{enc}(f_j(\mathbf{X}_j))$$

Then $\mathbf{X}_{f,1,2} \odot \mathbf{X}_{f,3,4} + \mathbf{X}_{f,5,6} \odot \mathbf{X}_{f,7,8}$ *is* $O(p^{-\delta})$*-close to a distribution with min-entropy* $(2+10\delta)\log p$*.*

Proof. Without loss of generality suppose f_1 has no fixed points. For all $i \in [3]$, using Lemma 11.2.12 we have that $\mathbf{X}_{f,2i-1,2i}$ is $O(p^{-\delta})$-close to a convex combination of at most 4 flat distributions A_{ij} of the form $(\mathrm{enc}(x_{2i-1}) + \mathrm{enc}(x_{2i}), \mathrm{enc}(f_1(x_{2i-1}) + \mathrm{enc}(f_2(x_{2i}))) : (x_{2i-1}, x_{2i}) \sim G_{ij}$ where $G_{ij} \subset \mathbb{F}_p^2$, $|G_{ij}| \geq p^{2-3\delta}$.

With probability $1 - O(p^{-\delta})$ over fixing of the sources $\mathbf{X}_7, \mathbf{X}_8$, we have $\mathbf{X}_{f,1,2} \odot \mathbf{X}_{f,3,4} + \mathbf{X}_{f,5,6} \odot x_{f,7,8}$ is $O(p^{-\delta})$-close to a convex combination of at most 4^3 distributions of the form $A_{1j_1} \cdot A_{2j_2} + \alpha \cdot A_{3j_3}$, $\alpha \in (\mathbb{F}_p^*)^4$. Since f_1 has no fixed points, by Lemma 11.2.15 with $\delta_1 = 3\delta$, we have that $A_{1j_1} \odot A_{2j_2} + \alpha \odot A_{3j_3}$ is $O(p^{-\delta})$-close to a distribution with min-entropy $(2+10\delta)\log p$. Hence, $\mathbf{X}_{f,1,2} \odot \mathbf{X}_{f,3,4} + \mathbf{X}_{f,5,6} \odot \mathbf{X}_{f,7,8}$ is $O(p^{-\delta})$-close to a distribution with min-entropy $(2 + 10\delta)\log p$.

□

11.2.5 Non-malleable extractors for functions with no fixed points

In this section we prove a special case of Theorem 3 where we have a restriction on the fixed points of the tampering functions. We use this result in the proof of Theorem 3.

Theorem 11.2.20. *There exists a constant* $\delta > 0$ *such that for every* n *there exists an explicit function* $\mathrm{nmExt} : (\{0,1\}^n)^8 \times \{0,1\}^{2n} \to \{0,1\}^m$*, such that if* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_8$ *are independent* $(n, (1-\delta)n)$*-sources,* $\mathbf{X}_9$ *an independent* $(2n, 2(1-\delta)n)$*-source and* $f_1, f_2, \ldots, f_9$ *are arbitrary tampering functions such that there exists* $j \in [8]$ *such that* f_j *has no fixed points, then*

$$|\mathrm{nmExt}(\mathbf{X}_1, \ldots, \mathbf{X}_9), \mathrm{nmExt}(f_1(\mathbf{X}_1), \ldots, f_9(\mathbf{X}_9)) - U_m, \mathrm{nmExt}(f_1(\mathbf{X}_1), \ldots, f_9(\mathbf{X}_9))| < 2^{-\Omega(n)}$$

Proof. We view each $\mathbf{X}_i$, $i \in [8]$, as a source on $\mathbb{F}_p$ for a prime p satisfying $2^n < p < 2^{n+1}$. If $p^{\tau_0} \leq 16$, we do a brute-force search for nmExt (in constant time). Thus assume $p^{\tau_0} > 16$.

Let $\mathrm{snmExt} : \{0,1\}^{2n} \times \{0,1\}^{2n} \to \{0,1\}^m$, $m = \Omega(n)$, be the seeded non malleable extractor

from Theorem 11.2.8. Define the functions

$$\mathrm{ext}_1(x_1, x_2, \ldots, x_8) = \sum_{i=0}^{1} \big(\mathrm{enc}(x_{4i+1}) + \mathrm{enc}(x_{4i+2})\big) \odot \big(\mathrm{enc}(x_{4i+3}) + \mathrm{enc}(x_{4i+4})\big)$$

$$\mathrm{nmExt}(x_1, \ldots, x_9) = \mathrm{snmExt}(\mathrm{ext}_1(x_1, \ldots, x_8), x_9)$$

We show that nmExt satisfies the conclusion of Theorem 11.2.20.

Let $S_i \subset \mathbb{F}_p$ be the support of the flat source $\mathbf{X}_i$ for all $i \in [8]$. Also let $S_9 \subset \{0,1\}^{2n}$ be the support of $\mathbf{X}_9$. We partition each S_i into S_{i0} and S_{i1} based on the pre-image of f_i as follows.

$$S_{i0} = \{s \in S_i : |f_i^{-1}(s) \cap S_i| \leq p^{20\delta}\}, \;\; S_{i1} = S_i \setminus S_{i1}.$$

Let $\mathbf{X}_{ij}$ be the flat source on S_{ij} for $j = 0, 1$.

We thus have

$$\begin{aligned} |\mathrm{nmExt}(\mathbf{X}_1, \ldots, \mathbf{X}_9), \mathrm{nmExt}(f_1(\mathbf{X}_1), \ldots, f_9(\mathbf{X}_9)) - U_m, \mathrm{nmExt}(f_1(\mathbf{X}_1), \ldots, f_9(\mathbf{X}_9))| \qquad (11.10) \\ \leq \sum_{I \in \{0,1\}^9} w_I \cdot |\mathrm{nmExt}(\mathbf{X}_{1I(1)}, \ldots, \mathbf{X}_{9I(9)}), \mathrm{nmExt}(f_1(\mathbf{X}_{1I(1)}), \ldots, f_9(\mathbf{X}_{9I(9)})) - \\ U_m, \mathrm{nmExt}(f_1(\mathbf{X}_{1I(1)}), \ldots, f_9(\mathbf{X}_{9I(9)}))| \qquad (11.11) \end{aligned}$$

where $w_I = \prod_{i=1}^{9} \left(\frac{|S_{iI(i)}}{|S_i|}\right)$.

We bound each term in (11.11). In particular we show that

$$\begin{aligned} w_I \cdot |\mathrm{nmExt}(\mathbf{X}_{1I(1)}, \ldots, \mathbf{X}_{9I(9)}), \mathrm{nmExt}(f_1(\mathbf{X}_{1I(1)}), \ldots, f_9(\mathbf{X}_{9I(9)})) - \\ U_m, \mathrm{nmExt}(f_1(\mathbf{X}_{1I(1)}), \ldots, f_9(\mathbf{X}_{9I(9)}))| < 2^{-\Omega(n)} \qquad (11.12) \end{aligned}$$

for each $I \in \{0,1\}^9$. Since there are $2^9 (= \textit{constant})$ such terms in (11.11), we get the required bound on (11.10).

We now prove (11.12). Fix any $I \in \{0,1\}^9$. The following two cases can occur.

1. Suppose for some $j \in [9]$, $|S_{jI(j)}| \leq p^{-\delta}|S_j|$. Then $w_I < p^{-\delta}$ and hence the bound in (11.12) follows.

2. Thus suppose $|S_{iI(i)}| \geq p^{-\delta}|S_j|$ for all $i \in [9]$.

   Define the random variables :

$$W^I = \text{ext}_1(\mathbf{X}_{1I(1)}, \ldots, \mathbf{X}_{8I(8)}), \quad V^I = \text{ext}_1(f_1(\mathbf{X}_{1I(1)}), \ldots, f_8(\mathbf{X}_{8I(8)}))$$

   We prove that the following holds.

$$\Pr_{v \sim \mathbf{V}^I}[(\mathbf{W}^I | \mathbf{V}^I = v) \text{ is } O(p^{-\delta})\text{-close to a distribution with min-entropy at least } 10\delta \log p] \geq 1 - p^{-\delta} \tag{11.13}$$

   The following two cases arise depending on I.

   (a) Suppose $I(j) = 0$ for all $j \in [8]$. It follows from Theorem 11.2.19 that $(\mathbf{W}^I, \mathbf{V}^I)$ is $p^{-\delta}$-close to a source with min-entropy $(2 + 20\delta) \log p$. Using Lemma 2.3.7, we have that

$$\Pr_{v \sim \mathbf{V}_i^I}[(\mathbf{W}_i^I | \mathbf{V}_i^I = v_i) \text{ is } O(p^{-\delta})\text{-close to a distribution with min-entropy at least } 10\delta \log p] \geq 1 - p^{-\delta}$$

   (b) Suppose there exists some $j \in [8]$ such that $I(j) = 1$. Consider fixing $f_j(\mathbf{X}_{jI(j)})$ and all $\mathbf{X}_{iI(i)}$, $i \in [8] \setminus \{j\}$. Without loss of generality suppose $j = 1$.
   Under this fixing W^I has min-entropy at least $20\delta \log p$ unless sources $\mathbf{X}_{3I(3)}, \mathbf{X}_{4I(4)}$ are fixed such that $\text{enc}(x_{3I(3)}) + \text{enc}(x_{4I(4)}) \notin (F_p^*)^2$. But it follows from Claim 11.2.14 that $\Pr[\text{enc}(\mathbf{X}_3) + \text{enc}(\mathbf{X}_4) \notin (\mathbb{F}_p^*)^2] < p^{-\delta}$. Thus,

$$\Pr_{v \sim V^I}[(W^I | V^I = v) \text{ is } O(p^{-\delta})\text{-close to a distribution with min-entropy at least } 20\delta \log p] = 1$$

   This completes the proof of (11.13).

We continue with the proof of (11.12). For each $i \in [C']$, define the set

$$Good^I = \{v \in \text{support}(\mathbf{V}^I) : (\mathbf{W}^I | \mathbf{V}^I = v) \text{ is } O(p^{-\delta})\text{-close to a distribution with min-entropy at least } 10\delta \log p\}$$

It follows from (11.13) that $\Pr_{v \sim \mathbf{V}^I}[v \in Good^I] > 1 - p^{-\delta}$.

It follows from Theorem 11.2.8 that snmExt is a strong 2-source extractor for independent sources on $2n$ bits with entropies k_1, k_2 respectively satisfying $k_1 + k_2 \geq (2+\delta)n$.

Thus we have,

$$\begin{aligned} &|\text{snmExt}(W^I, \mathbf{X}_{9I(9)}), V^I, \mathbf{X}_{9I(9)} - U_m, V^I, \mathbf{X}_{9I(9)}| \\ &\leq (\Pr[\mathbf{V}^I \notin Good^I]) + 2^{-\Omega(n)} + p^{-\delta} \leq 2p^{-\delta} + 2^{-\Omega(n)} = 2^{-\Omega(n)} \end{aligned}$$

Since $\text{nmExt}(f_1(\mathbf{X}_{1I(1)}), \ldots, f_9(\mathbf{X}_{9,I(9)}))$ is a deterministic function of the random variables V^I and $\mathbf{X}_{9I(9)}$, the bound in (11.12) is now immediate.

□

11.2.6 Non-malleable extractor for arbitrary functions

We now prove a slightly stronger version of Theorem 28.

Theorem 11.2.21 (Theorem 28 restated, stronger version)**.** *There exists a constant $\delta > 0$ such that for every n there exists an explicit function* $\text{nmExt} : (\{0,1\}^n)^8 \times \{0,1\}^{2n} \to \{0,1\}^m$, *$m = \Omega(n)$, such that if $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_8$ are independent $(n, (1-\delta)n)$-sources, $\mathbf{X}_9$ an independent $(2n, 2(1-\delta)n)$-source and $f_1, f_2, \ldots, f_9$ are arbitrary tampering functions, such that at least one of the f_i's have no fixed points. Then*

$$|\text{nmExt}(\mathbf{X}_1, \ldots, \mathbf{X}_9), \text{nmExt}(f_1(\mathbf{X}_1), \ldots, f_9(\mathbf{X}_9)) - \mathbf{U}_m, \text{nmExt}(f_1(\mathbf{X}_1), \ldots, f_9(\mathbf{X}_9))| \leq 2^{-\Omega(n)}.$$

Proof. We view each $\mathbf{X}_i$, $i \in [8]$, as a source on $\mathbb{F}_p$ for a prime p satisfying $2^n < p < 2^{n+1}$. We assume $p^{\tau_0} > 16$ (else we do a constant time brute-force search for nmExt).

Let $\text{snmExt} : \{0,1\}^{2n} \times \{0,1\}^{2n} \to \{0,1\}^m$, $m = \Omega(n)$, be the seeded non-malleable extractor from Theorem 11.2.8.

Define the functions

$$\text{ext}_1(x_1, x_2, \ldots, x_8) = \sum_{i=0}^{1} \big(\text{enc}(x_{4i+1}) + \text{enc}(x_{4i+2})\big) \odot \big(\text{enc}(x_{4i+3}) + \text{enc}(x_{4i+4})\big)$$

$$\text{nmExt}(x_1, \ldots, x_9) = \text{snmExt}(\text{ext}_1(x_1, \ldots, x_8), x_9)$$

We need the following claims.

Claim 11.2.22. *Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_8$ be sources on $\mathbb{F}_p$ with min-entropy $(1-2\delta)\cdot\log p$. Then $\text{ext}_1(\mathbf{Y}_1, \ldots, \mathbf{Y}_8)$ is $2^{-\Omega(n)}$-close to a source with min-entropy $(1-2\delta)\cdot 2\log p$.*

Proof. We claim that $\text{enc}(\mathbf{Y}_1) + \text{enc}(\mathbf{Y}_2)$ is a source with min-entropy $2(1-2\delta)\log p - 2$. This follows from the fact that $(y_1 + y_2, y_1^4 + y_1^2 + y_1 + y_2 + y_2^2 + y_2) = (a, b)$ has at most 4 solutions in (y_1, y_2). Also it follows from Claim 11.2.14 that $\Pr[\text{enc}(\mathbf{Y}_3) + \text{enc}(\mathbf{Y}_4) \notin (\mathbb{F}_p^*)^2] < p^{-\delta}$. Thus $\text{ext}_1(\mathbf{Y}_1, \ldots, \mathbf{Y}_8)$ is $p^{-\delta}$-close to a source with min-entropy $2(1-2\delta)\log p - 2$. □

Claim 11.2.23. *Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_8$ be independent $(n, (1-2\delta)n)$-sources and $\mathbf{Y}_9$ an independent $(2n, 2(1-2\delta)n)$-source . Then*

$$|\text{nmExt}(\mathbf{Y}_1, \ldots, \mathbf{Y}_9) - U_m| < 2^{-\Omega(n)}$$

Proof. Follows directly from Claim 11.2.22 and Theorem 11.2.8. □

For each $i \in [8]$, let $S_i \subset \mathbb{F}_p$ be the support of the (flat) source $\mathbf{X}_i$. Let $S_9 \subset \{0,1\}^{2n}$ be the support of $\mathbf{X}_9$. We partition each S_i into S_{i0} and S_{i1} such that f_i has no fixed points in S_{i1}. Thus

$$S_{i0} = \{s \in S_i : f_i(s) = s\}, S_{i1} = S_i \setminus S_{i0}$$

Let $\mathbf{X}_{ij}$ be the flat source that is supported on S_{ij}, $i = 1, .., 9$, $j = 0, 1$. Let f_i^I denote f_i with its domain restricted to the set $S_{iI(i)}$. Thus f_i^I is a function from $S_{iI(i)}$ to $\mathbb{F}_p$.

For any 0-1 vector I, let $I(i)$ denote the i'th co-ordinate in I. Let $w_I = \prod_{i=1}^{9} \frac{|S_{iI(i)}|}{|S_i|}$ for $I \in \{0,1\}^9$.

Recall that to prove Theorem 11.2.21, we need to show the following bound.

$$|\mathrm{nmExt}(\mathbf{X}_1, \ldots, \mathbf{X}_9), \mathrm{nmExt}(f_1(\mathbf{X}_1), \ldots, f_9(\mathbf{X}_9)) - \mathbf{U}_m, \mathrm{nmExt}(f_1(\mathbf{X}_1), \ldots, f_9(\mathbf{X}_9))| < 2^{-\Omega(n)} \tag{11.14}$$

We have

$$(11.14) \leq \sum_{I \in \{0,1\}^9 \setminus \{\vec{0}\}} w_I \cdot |\mathrm{nmExt}(\mathbf{X}_{1I(1)}, \ldots, \mathbf{X}_{9I(9)}), \mathrm{nmExt}(f_1^I(\mathbf{X}_{1I(1)}), \ldots, f_9^I(\mathbf{X}_{9I(9)})) - \mathbf{U}_m, \mathrm{nmExt}(f_1^I(\mathbf{X}_{1I(1)}), \ldots, f_9^I(\mathbf{X}_{9I(9)}))| \tag{11.15}$$

We prove the following claim.

Claim 11.2.24. *For every $I \in \{0,1\}^9 \setminus \{\vec{0}\}$ the following holds:*

$$w_I \cdot |\mathrm{nmExt}(\mathbf{X}_{1I(1)}, \ldots, \mathbf{X}_{9I(9)}), \mathrm{nmExt}(f_1(\mathbf{X}_{1I(1)}), \ldots, f_9(\mathbf{X}_{9I(9)})) - U_m, \mathrm{nmExt}(f_1^I(\mathbf{X}_{1I(1)}), \ldots, f_9^I(\mathbf{X}_{9I(9)}))| < 2^{-\Omega(n)} \tag{11.16}$$

We use the above claim to conclude (11.14).

Proof of (11.14) using Claim 11.2.24. Note that there are $2^9 - 1$ terms in RHS of (11.15). Each term is bounded by $2^{-\Omega(n)}$ using Claim 11.2.24. We can thus bound LHS of (11.14) by $2^{-\Omega(n)}$. □

Proof of Claim 11.2.24. Fix some $I \in \{0,1\}^9 \setminus \{\vec{0}\}$.

We split the proof into the following cases.

1. If for some $i \in [9]$, $|S_{iI(i)}| < p^{-\delta}|S_i|$, then $w_I < p^{-\delta}$ and hence the bound in (11.16) follows.

2. Thus suppose $|S_{iI(i)}| \geq p^{-\delta}|S_i|$ for all $i \in [9]$. We consider the following cases.

   (a) Suppose there exists some $j \in [8]$ such that $I(j) = 1$. In this case we use Theorem 11.2.20 to conclude the bound in (11.16).

   (b) Suppose for all $i \in [8]$, $I(i) = 0$. We note that $I(9) = 1$ since $I \neq \vec{0}$. Thus all f_i^I, $i \in [8]$, are the identity functions over their respective domains and f_9^I has no fixed points.

   Using Claim 11.2.22, we have $\mathrm{ext}_1(\mathbf{X}_{1I(1)}, \ldots, \mathbf{X}_{8I(8)})$ is $2^{-\Omega(n)}$-close to a source $\mathbf{Z}$ with min-entropy $(1-2\delta)\cdot 2n$.

   Define the random variable: $\mathbf{W}^I = \mathrm{ext}_1(\mathbf{X}_{1I(1)}, \ldots, \mathbf{X}_{8I(8)})$.

   Thus we have

$$\begin{aligned}
&|\mathrm{nmExt}(\mathbf{X}_{1I(1)}, \ldots, \mathbf{X}_{9I(9)}), \mathrm{nmExt}(f_1(\mathbf{X}_{1I(1)}), \ldots, f_9(\mathbf{X}_{9I(9)}))\\
&\qquad -\mathbf{U}_m, \mathrm{nmExt}(f_1(\mathbf{X}_{1I(1)}), \ldots, f_9(\mathbf{X}_{9I(9)}))|\\
&= |\mathrm{snmExt}(W^I, \mathbf{X}_{9I(9)}), \mathrm{snmExt}(W^I, f_9^I(\mathbf{X}_{9I(9)})) - \mathbf{U}_m, \mathrm{snmExt}(W^I, f_9^I(\mathbf{X}_{9I(9)}))|\\
&\leq |\mathrm{snmExt}(\mathbf{Z}, \mathbf{X}_{9I(9)}), \mathrm{snmExt}(\mathbf{Z}, f_9^I(\mathbf{X}_{9I(9)})) - \mathbf{U}_m, \mathrm{snmExt}(\mathbf{Z}, f_9^I(\mathbf{X}_{9I(9)}))| + 2^{-\Omega(n)}
\end{aligned}$$

   Note that $\mathbf{Z}$ and $\mathbf{X}_{9I(9)}$ are independent sources on $\{0,1\}^{2n}$, each with min-entropy rate $> (1-2\delta)$ and f_9^I has no fixed points. Thus by Theorem 11.2.8, we have

$$|\mathrm{snmExt}(\mathbf{Z}, \mathbf{X}_{9I(9)}), \mathrm{snmExt}(\mathbf{Z}, f_9^I(\mathbf{X}_{9I(9)})) - U_m, \mathrm{snmExt}(\mathbf{Z}, f_9^I(\mathbf{X}_{9I(9)}))| \leq 2^{-\Omega(n)}$$

   Thus, the bound in (11.16) follows.

This completes the proof of Claim 11.2.24. □

□

11.2.7 Proof of the sum-product estimate over $\mathbb{F}_p^4$

We closely follow the proof of the sum-product estimate by Bourgain in [Bou05a] and prove Theorem 11.2.10, which we restate.

Theorem 1.9. *There exists $\tau_0 > \tau_1 > 0$ such that the following holds: Let A be a subset of $\mathbb{F}_p^4$ satisfying $|A \cap (\mathbb{F}_p^*)^4| \geq \frac{|A|}{2}$. Suppose that for any subset $A_1 \subseteq A$ satisfying $|A_1| \geq p^{-\tau_1}|A|$, the following conditions holds.*

1. *$\Pi_{\{1,2\}}(A_1) \geq p^{1+\tau_0}$ and $\Pi_{\{3,4\}}(A_1) \geq p^{1+\tau_0}$.*
2. *$A_1 \nsubseteq P$, where P is a 2-dimensional linear subspace of $\mathbb{F}_p^4$ of form*
   (a) *$\{(x_1, x_2, c_1x_1, c_2x_2) : x_1 \in \mathbb{F}_p, x_2 \in \mathbb{F}_p\}$ or*
   (b) *$\{(x_1, x_2, c_2x_2, c_1x_1) : x_1 \in \mathbb{F}_p, x_2 \in \mathbb{F}_p\}$.*

Then there exists some constant $\tau > 0$ (depending on τ_0, τ_1) such that if $|A| < p^{7/3-\tau_1}$, then

$$|A + A| + |A \odot A| > p^{\tau}|A|$$

We introduce some notations.

Definition 11.2.25. *Let $S \subseteq \mathbb{F}_p^n$ be any set of vectors. Define $S^{\odot 2} = S \odot S$ and $S^{\odot(k+1)} = S^{\odot k} \odot S$ for $k \geq 2$.*

We prove Theorem 11.2.10 using the following lemmas.

Lemma 11.2.26. *Let B be any subset of $\mathbb{F}_p^4$ such that $|\Pi_{\{1,2\}}(B)| \geq p^{1+\tau_0}$ and $|\Pi_{\{3,4\}}(B)| \geq p^{1+\tau_0}$. Then one of the following holds.*

1. *There exists constant $k = k(\tau_0)$ such that $|kB^{\odot k}| \geq p^{7/3}$ or*
2. *$B \subseteq P$ where P is a 2-dimensional linear subspace of $\mathbb{F}_p^4$ of the form*

(a) $\{(x_1, x_2, c_1x_1, c_2x_2) : x_1 \in \mathbb{F}_p, x_2 \in \mathbb{F}_p\}$ *or*

(b) $\{(x_1, x_2, c_2x_2, c_1x_1) : x_1 \in \mathbb{F}_p, x_2 \in \mathbb{F}_p\}$.

Lemma 11.2.27. *Let* $B \subset (\mathbb{F}_p^*)^4$ *such that* $|B| \geq p^{1+\tau_0}$ *and* $|B+B| + |B \odot B| \leq p^{\tau}|B|$. *Fix any* $k > 0$. *Then, there is a subset* B_1 *of* B *such that*

1. $|B_1| \geq p^{-\tau_1}|B|$ *and*
2. $|kB_1^{\odot k}| \leq p^{\tau_1}|B_1|$

where $\tau_1 = p^{3k^2}\tau$.

Proof of Theorem 11.2.10. We replace A with its intersection with $(\mathbb{F}_p^*)^4$. Choose τ small enough such for $k = k(\tau_0)$ (where $k(\tau_0)$ is the constant from Lemma 11.2.26), it holds that: $p^{3k^2}\tau < \tau_1$. Suppose that $|A+A| + |A \odot A| \leq p^{\tau}|A|$. Using Lemma 11.2.27, there exists a subset A_1, $|A_1| \geq p^{-\tau_1}|A|$, such that $|kA_1^{\odot k}| \leq |A|p^{\tau_1}$. Further, we have that A_1 satisfies the hypothesis of Lemma 11.2.26. Suppose, conclusion (1) of Lemma 11.2.26 holds. This implies that $|A| \geq p^{7/3-\tau_1}$ which contradicts our assumption on the size of A. Further, from the assumptions on the structure of A_1, we see that conclusion (2) in Lemma 11.2.26 cannot hold. Thus, it must be that $|A+A| + |A \odot A| > p^{\tau}|A|$. □

Lemma 11.2.27 follows directly from Lemma 4 in [Bou05a] by noticing that their proof works over $(\mathbb{F}_p^*)^4$ as well. Hence we do not present the proof of Lemma 11.2.27.

Thus we focus on proving Lemma 11.2.26.

We require the following lemma which was proved by Bourgain [Bou05a].

Lemma 11.2.28. *For any* $B \subseteq \mathbb{F}_p^2$ *such that* $|B| \geq p^{1+\tau_0}$ *there exists a constant* $k = k(\tau_0)$ *such that* $|kB^{\odot k}| = p^2$.

We now proceed to prove Lemma 11.2.26.

Proof of Lemma 11.2.26. Let B_{ij} denote $\Pi_{\{i,j\}}(B)$. Using Lemma 11.2.28, there exists some k_0 such that $|kB_{12}^{\odot k}| = p^2$, $|kB_{34}^{\odot k}| = p^2$ for $k \geq k_0$. We split the proof into two cases.

1. Suppose there exists some $k \geq k_0$ such that $|kB^{\odot k}| > p^2$.

   Thus, it must be the case that the projection map $\Pi_{\{1,2\}}$ is not one-one on $kB^{\odot k}$. Thus there exists $b, b' \in kB^{\odot k}$ such that $\Pi_{\{1,2\}}(b) = \Pi_{\{1,2\}}(b')$ but $\Pi_{\{3,4\}}(b) \neq \Pi_{\{3,4\}}(b')$. Consider the set

   $$kB^{\odot k} - (b-b')kB^{\odot k} = \{\big(x_1, x_2, x_3 - (b_3 - b_3')y_3, x_4 - (b_4 - b_4')y_4\big) : \\ (x_1, x_2, x_3, x_4) \in kB^{\odot k}, (y_1, y_2, y_3, y_4) \in kB^{\odot k}\}$$

   Notice that (x_1, x_2) takes all values of $\mathbb{F}_p^2$ since $|\Pi_{\{1,2\}}(kB^{\odot k})| = p^2$. Similarly (y_3, y_4) takes all values of $\mathbb{F}_p \times \mathbb{F}_p$ since $|\Pi_{\{3,4\}}(kB^{\odot k})| = p^2$. Further, at least one of $(b_3 - b_3')$ or $(b_4 - b_4')$ is non zero. Without loss of generality, suppose $b_3 - b_3' \neq 0$. Then, for any fixing of $x \in kB^{\odot k}$, $\Pi_3(x - (b-b')kB^{\odot k}) = \mathbb{F}_p$ and hence $|kB^{\odot k} - (b-b')kB^{\odot k}| \geq p^3$.

   We observe that

   $$kB^{\odot k} - (b-b')kB^{\odot k} \subseteq kB^{\odot k} - (kB^{\odot k} - kB^{\odot k})kB^{\odot k} \subseteq k'B^{\odot k'} - k'B^{\odot k'}$$

   , where $k' = 3k^2$. Using Lemma 11.2.5 with $A = k'B^{\odot k'}$ and recalling that $|k'B^{\odot k'}| > p^2$, we have

   $$\begin{aligned} |k'B^{\odot k'} + k'B^{\odot k'}| &\geq \left(|k'B^{\odot k'} - k'B^{\odot k'}||k'B^{\odot k'}|^2\right)^{1/3} \\ &> (p^3p^4)^{1/3} = p^{7/3} \end{aligned}$$

   Setting a new $k = 2k'$, we have $|kB^{\odot k}| \geq p^{7/3}$.

2. Suppose $|kB^{\odot k}| = p^2$ for all $k \geq k_0$. Thus in particular we have

$$|k_0B^{\odot k_0} + k_0B^{\odot k_0}| = |k_0B^{\odot k_0}|$$

and

$$|k_0B^{\odot k_0} \odot k_0B^{\odot k_0}| = |k_0B^{\odot k_0}|$$

Thus $k_0B^{\odot k_0}$ must be a 2-dimensional affine subspace of $\mathbb{F}_p^4$.

Let $k_0B^{\odot k_0} = \{z + \lambda v + \mu w : \lambda, \mu \in \mathbb{F}_p\}$, $z, v, w \in \mathbb{F}_p^4$. To complete the argument, we prove the following claims about the structure of z, v, w.

Claim 11.2.29. *We can assume $v = (1, 0, \alpha_1, \alpha_2)$ and $w = (0, 1, \beta_1, \beta_2)$ such that $span\{(\alpha_1, \alpha_2), (\beta_1, \beta_2)\} = \mathbb{F}_p^2$*

Proof. The proof follows from the observation that $\Pi_{\{1,2\}}(k_0B^{\odot k_0}) = \Pi_{\{3,4\}}(k_0B^{\odot k_0}) = \mathbb{F}_p^2$. □

Claim 11.2.30. *Let $v = (1, 0, \alpha_1, \alpha_2)$ and $w = (0, 1, \beta_1, \beta_2)$. Then $\alpha_i\beta_i = 0$ for $i \in [2]$. Further $z = 0$.*

We show how to complete the proof of Lemma 11.2.26, before proving the above claim.

Proof of Lemma 11.2.26 using Claim 11.2.29 and Claim 11.2.30. Since we have $\alpha_1\beta_1 = 0$, suppose $\alpha_1 = 0$. It follows from Claim 11.2.29 and Claim 11.2.30 that $\beta_1 \neq 0$, $\alpha_2 \neq 0$ and $\beta_2 = 0$.

Thus $k_0B^{\odot k_0} = \{z + \lambda v + \mu w : \lambda, \mu \in \mathbb{F}_p\} = \{(\lambda, \mu, \beta_1\mu, \alpha_2\lambda) : \lambda, \mu \in \mathbb{F}_p\}$.

Fix any $y = (y_1, y_2, y_3, y_4) \in k_0B^{\odot(k_0-1)} \cap (\mathbb{F}_p^*)^4$. Note that there exists such a y since $B \cap (\mathbb{F}_p^*)^4 \neq \emptyset$ and $k_0B^{\odot k_0} \cap (\mathbb{F}_p^*)^4 \neq \emptyset$.

For any $x = (x_1, x_2, x_3, x_4) \in B$, since $x \odot y \in k_0 B^{\odot k_0} = \{(\lambda, \mu, \beta_1\mu, \alpha_2\lambda) : \lambda, \mu \in \mathbb{F}_p\}$, there exists λ, μ such that the following relations hold :

$$x_4 = y_4^{-1}\alpha_2 x_1 y_1, \quad x_3 = y_3^{-1}\beta_1 x_2 y_2$$

Thus

$$B \subseteq \{(x_1, x_2, c_2x_2, c_1x_1) : x_1, x_2 \in \mathbb{F}_p\}$$

where $c_1 = y_4^{-1}\alpha_2 y_1, c_2 = y_3^{-1}\beta_1 y_2$.

For the case when $\alpha_1 \neq 0$ (and hence $\beta_1 = 0$), we use an identical argument to derive that $B \subseteq \{(x_1, x_2, c_1x_1, c_2x_2 : x_1, x_2) \in \mathbb{F}_p\}$. □

We conclude by proving Claim 11.2.30.

Proof of Claim 11.2.30. Let $S = (k_0 B^{\odot k_0}) \odot (k_0 B^{\odot k_0})$. Recall that $k_0 B^{\odot k_0} = \{z + \lambda v + \mu w : \lambda, \mu \in \mathbb{F}_p\}$ where $v = (1, 0, \alpha_1, \alpha_2), w = (0, 1, \beta_1, \beta_2)$ and $|S| = |k_0 B^{\odot k_0}| = p^2$. Thus for each $i \in [4]$,

$$\Pi_i(S) = \{\pi_i(\lambda_1, \lambda_2, \mu_1, \mu_2) : \lambda_1, \lambda_2, \mu_1, \mu_2 \in \mathbb{F}_p\}$$

where

$$\begin{aligned}
\pi_1(\lambda_1,\lambda_2,\mu_1,\mu_2) &= \pi_1(\lambda_1,\lambda_2) = (z_1+\lambda_1)(z_1+\lambda_2)\\
&= \lambda_1\lambda_2 + (\lambda_1+\lambda_2)z_1 + z_1^2\\
\pi_2(\lambda_1,\lambda_2,\mu_1,\mu 2) &= \pi_2(\mu_1,\mu_2) = (z_2+\mu_1)(z_2+\mu_2)\\
&= \mu_1\mu_2 + (\mu_1+\mu_2)z_2 + z_2^2\\
\pi_3(\lambda_1,\lambda_2,\mu_1,\mu_2) &= (\lambda_1\alpha_1+\mu_1\beta_1+z_3)(\lambda_2\alpha_1+\mu_2\beta_1+z_3)\\
&= \lambda_1\lambda_2\alpha_1^2 + \mu_1\mu_2\beta_1^2 + \alpha_1\beta_1(\lambda_1\mu_2+\lambda_2\mu_1)+\\
&\qquad (\lambda_1+\lambda_2)\alpha_1 z_3 + (\mu_1+\mu_2)\beta_1 z_3 + z_3^2\\
\pi_4(\lambda_1,\lambda_2,\mu_1,\mu_2) &= (\lambda_1\alpha_2+\mu_1\beta_2+z_4)(\lambda_2\alpha_2+\mu_2\beta 2+z_4)\\
&= \lambda_1\lambda_2\alpha_2^2 + \mu_1\mu_2\beta_2^2 + \alpha_2\beta_2(\lambda_1\mu_2+\lambda_2\mu_1)+\\
&\qquad (\lambda_1+\lambda_2)\alpha_2 z_4 + (\mu_1+\mu_2)\beta_2 z_4 + z_4^2
\end{aligned}$$

- We prove $\alpha_i\beta_i = 0$, for $i = 1, 2$. Suppose not. Let $\alpha_1\beta_1 \neq 0$.

  Fix $\lambda_2 = a_2 \neq -z_1$ and let $\lambda_1 = a_1 \neq \lambda_2$ and let $\pi_1(a_1, a_2) = a$. Note that $\pi_1(a_1, a_2) = \pi(b_1, a_2)$ iff $a_1 = b_1$. Thus $|\{\pi_1(x, a_2) : x \in \mathbb{F}_p \setminus \{a_2\}\}| = p - 1$.

  We claim that for any such fixing of $\lambda_1 = a_1, \lambda_2 = a_2$, there exists μ_1, μ_2 such that $\pi_2(\mu_1, \mu_2) = b$ and $\pi_3(a_1, a_2, \mu_1, \mu_2) = c$ for at least $O(p^2)$ pairs $(b, c) \in \mathbb{F}_p^2$. Suppose

$$\begin{aligned}
\pi_2(\mu_1,\mu_2) &= \mu_1\mu_2 + (\mu_1+\mu_2)z_2 + z_2^2 = b\\
\pi_3(a_1,a_2,\mu_1,\mu_2) &= \beta_1^2\mu_1\mu_2 + \gamma_1\mu_1 + \gamma_2\mu_2 + \gamma_3 = c
\end{aligned}$$

  where $\gamma_1, \gamma_2, \gamma_3 \in \mathbb{F}_p$ are constants (does not depend on μ_1, μ_2) . By our choice of λ_1, λ_2, we have that $\gamma_1 \neq \gamma_2$ and hence the above system of equations has at most two pairs of values of (μ_1, μ_2) which satisfy it. Since (μ_1, μ_2) takes p^2 values, there at least $p^2/2$ distinct pairs (b, c) such that there $(\pi_2(\mu_1, \mu_2), \pi_3(\lambda_1, \lambda_2, \mu_1, \mu_2)) = (b, c)$.

  Thus we have shown that there exists $\lambda_1, \lambda_2, \mu_1, \mu_2$ such that $(\pi_1(\lambda_1, \lambda_2), \pi_2(\mu_1, \mu_2), \pi_3(\lambda_1, \lambda_2, \mu_1, \mu_2)) =$

(a, b, c) for at least $\frac{1}{2}(p-1)p^2$ distinct tuples $(a, b, c) \in \mathbb{F}_p^3$, which is a contradiction since $|S| = p^2$. Thus $\alpha_1\beta_1 = 0$. A similar argument implies that $\alpha_2\beta_2 = 0$.

- We now prove $z = 0$. Suppose $\alpha_1 = 0$. Thus $\beta_1 \neq 0$, $\alpha_2 \neq 0$ and $\beta_2 = 0$. We again fix $\lambda_2 = a_2 \neq -z_1$ and let $\lambda_1 = a_1 \neq \lambda_2$. Let $(b, c) \in \mathbb{F}_p^2$. We bound the number of (μ_1, μ_2) such that $(\pi_2(\mu 1, \mu_2), \pi_3(a_1, a_2, \mu_1, \mu_2) = (b, c)$. We have the following equations.

$$\mu_1\mu_2 + (\mu_1 + \mu_2)z_2 + z_2^2 = b$$
$$\beta_1^2\mu_1\mu_2 + \beta_1 z_3(\mu_1 + \mu_2) + \gamma_0 = c$$

We see that the number of solutions of the above pair of equations is bounded by 2 unless $z_3 = \beta_1 z_1$. It follows that if $z_3 \neq \beta_1 z_2$, there exists $(\lambda_1, \lambda_2, \mu_1, \mu_2)$ such that $(\pi_1(\lambda_1, \lambda_2), \pi_2(\mu_1, \mu_2), \pi_3(\lambda_1, \lambda_2, \mu_1, \mu_2)) = (a, b, c)$ for at least $\frac{1}{2}(p-1)p^2$ distinct tuples $(a, b, c) \in \mathbb{F}_p^3$, which is a contradiction. Thus suppose $z_3 = \beta_1 z_2$.

Using an identical argument (but now fixing μ_1, μ_2 appropriately in π_2 and arguing about the range of π_1 and π_4 upon varying λ_1, λ_2), we get that $z_4 = \alpha_2 z_1$. Thus $z = (z_1, z_2, \beta_1 z_2, \alpha_2 z_1) = z_1 \cdot (1, 0, 0, \alpha_2) + z_2 \cdot (0, 1, \beta_1, 0) = z_1 \cdot v + z_2 \cdot w \in span\{v, w\}$. Hence we can take $z = 0$. □

□

11.3 An Explict Seedless $(2, t)$-Non-Malleable Extractor Construction

The result in this section is based on [CGL16]. The extractor construction is similar to the seeded t-non-malleable in Section 4.3. Thus, we present the construction and omit the proof.

Subroutines and Parameters

1. Let γ be a small enough constant and C a large one. Let $t = n^{\gamma/C}$.

2. Let $n_1 = n^{\beta_1}$, $\beta_1 = 10\gamma$. Let IP : $\{0,1\}^{n_1} \times \{0,1\}^{n_1} \to \{0,1\}^{n_2}$, $n_2 = \frac{n_1}{10}$, be the strong two-source extractor from Theorem 2.5.3.

3. Let $\mathcal{C}$ be an explicit $[\frac{n}{\alpha}, n, \frac{1}{10}]$-binary linear error correcting code with encoder $E : \{0,1\}^n \to \{0,1\}^{\frac{n}{\alpha}}$. Such explicit codes are known, for example from the work of Alon et al. [ABN+92].

4. Let Samp : $\{0,1\}^{n_2} \to \left[\frac{n}{\alpha}\right]$ be the sampler from Theorem 2.4.2. Let the number of samples $t_{\text{Samp}} = n^{\beta_2}$. Thus, $\beta_2 \leq \beta_1$.

5. Let $\ell = 2(n^{\beta_1} + n^{\beta_2})$. Thus $\ell \leq n^{11\gamma}$.

6. We set up the parameters for the components used by flip-flop (computed by Algorithm 1) as follows.

   (a) Let $n_3 = n^{\beta_3}, n_4 = n^{\beta_4}$, with $\beta_3 = 100\gamma$ and $\beta_4 = 50\gamma$.

   Let $\text{Ext}_q : \{0,1\}^{n_3} \times \{0,1\}^{n_4} \to \{0,1\}^{n_4}$ be the strong seeded linear extractor from Theorem 2.1.5 set to extract from min-entropy $k_q = \frac{n_3}{4}$ with error $\epsilon = 2^{-\Omega(n^{\gamma_q})}$, $\gamma_q = \frac{\beta_4}{2}$. Thus, by Theorem 2.1.5, we have that the seed length $d_q = O\left(\frac{\log^2(n_3/\epsilon)}{\log(k_q/n_4)}\right) = O(n^{2\gamma_q}) = n_4$.

   Let $\text{Ext}_w : \{0,1\}^n \times \{0,1\}^{n_4} \to \{0,1\}^{n_4}$ be the strong linear seeded extractor from Theorem 2.1.5 set to extract from min-entropy $k_w = \frac{n}{2}$ with error $\epsilon = 2^{-\Omega(n^{\gamma_q})}$.

   (b) Let laExt : $\{0,1\}^n \times \{0,1\}^{n_3} \to \{0,1\}^{2n_4}$ be the look ahead extractor used by 2laExt (recall that the parameters in the alternating extraction protocol are set as $m = n_4, u = 2$ where u is the number of steps in the protocol, m is the length of each random variable that is communicated between the players, and $\text{Ext}_q, \text{Ext}_w$ are the strong seeded extractors used in the protocol.).

   (c) Let Ext : $\{0,1\}^n \times \{0,1\}^{n_4} \to \{0,1\}^{n_3}$ be the linear strong seeded extractor from Theorem 2.1.5 set to extract from min-entropy $\frac{n}{2}$ with seed length n_4 and error $2^{-\Omega(n^{\beta_4/2})}$.

7. Let ACB be the function computed by Algorithm 9, which uses the function flip-flop set up as above.

Algorithm 10: nmExt(x,y)

Input: Bit strings x, y, each of length n.
Output: A bit string of length n_4.

1 Let $x_1 = \text{Slice}(x, n_1)$, $y_1 = \text{Slice}(y, n_1)$. Compute $v = \text{IP}(x, y)$.
2 Compute $T = \text{Samp}(v) \subset [\frac{n}{\alpha}]$.
3 Let $z = x_1 \circ x_2 \circ y_1 \circ y_2$ where $x_2 = (E(x))_{\{T\}}, y_2 = (E(y))_{\{T\}}$.
4 Output $\text{ACB}(x, y, z)$.

By following along the lines of the proof of Theorem 4.3.1, it is easy to obtain the following result.

Theorem 11.3.1. *Let* nmExt *be the function computed by Algorithm* 3. *Then* nmExt *is a seedless* $(2, t)$*-non-malleable extractor with error* $2^{-n^{\Omega(1)}}$.

Chapter 12

Non-Malleable Codes

[1] Error-correcting codes encode a message m into a longer codeword c enabling recovery of m even after part of c is corrupted. We can view this corruption as a tampering function f acting on the codeword, where f is from some small allowable family $\mathcal{F}$ of tampering functions. The strict requirement of retrieving the encoded message m imposes restrictions on the kind of tampering functions that can be handled. Unique decoding is limited by the minimum distance of the codeword, and various bounds are known in the case of list decoding. Hence, many natural classes of tampering functions cannot be handled in this framework.

One might hope to achieve a weaker goal of only detecting errors, possibly with high probability. Cramer et al. [CDF+08] constructed one such class of error-detecting codes, known as Algebraic Manipulation Detection codes (AMD codes), where the allowable tampering functions consist of all functions of the form $f_a(x) = a+x$. However error detection is impossible with respect to the family of constant functions. This follows since one cannot hope to detect errors against a function that always outputs some fixed codeword.

Dziembowski, Pietrzak and Wichs [DPW10] introduced non-malleable codes as a natural generalization of error-detecting codes. Informally, a non-malleable code with respect to a tam-

[1] parts of this chapter have been previously published [CZ14, CGL16]

pering function family $\mathcal{F}$ is equipped with a randomized encoder Enc and a deterministic decoder Dec such that $\mathrm{Dec}(\mathrm{Enc}(m)) = m$ and for any tampering function $f \in \mathcal{F}$ the following holds: for any message m, $\mathrm{Dec}(f(\mathrm{Enc}(m)))$ is either the message m or is ϵ-close (in statistical distance) to a distribution D_f independent of m. The parameter ϵ is called the error.

We now formally define non-malleable codes. We need to define the following function.

$$\mathrm{copy}(x, y) = \begin{cases} x & \text{if } x \neq same^\star \\ y & \text{if } x = same^\star \end{cases}$$

$$\mathrm{copy}^{(t)}((x_1, \ldots, x_t), (y_1, \ldots, y_t)) = (\mathrm{copy}(x_1, y_1), \ldots, \mathrm{copy}(x_t, y_t))$$

Definition 12.0.1 (Coding schemes)**.** *Let* $\mathrm{Enc} : \{0,1\}^k \to \{0,1\}^n$ *and* $\mathrm{Dec} : \{0,1\}^n \to \{0,1\}^k \cup \{\perp\}$ *be functions such that* Enc *is a randomized function (i.e. it has access to a private randomness) and* Dec *is a deterministic function. We say that* $(\mathrm{Enc}, \mathrm{Dec})$ *is a coding scheme with block length* n *and message length* k *if for all* $s \in \{0,1\}^k$, $\Pr[\mathrm{Dec}(Enc(s)) = s] = 1$ *(the probability is over the randomness in* Enc*).*

Let $\mathcal{F}_n$ denote the set of all functions mapping n-bit strings to n-bit strings.

Definition 12.0.2 (Non-malleable codes)**.** *A coding scheme* $(\mathrm{Enc}, \mathrm{Dec})$ *with block length* n *and message length* k *is a non-malleable code with respect to a family of tampering functions* $\mathcal{F} \subset \mathcal{F}_n$ *and error* ϵ *if for every* $f \in \mathcal{F}$ *there exists a random variable* D_f *on* $\{0,1\}^k \cup \{same^\star\}$ *which is independent of the randomness in* Enc *such that for all messages* $s \in \{0,1\}^k$*, it holds that*

$$|\mathrm{Dec}(f(\mathrm{Enc}(s))) - \mathrm{copy}(D_f, s)| \leq \epsilon$$

The rate of a non-malleable code $\mathcal{C}$ is given by $\frac{k}{n}$.

As an easy example, suppose the tampering function family at hand is $\mathcal{F}_{constant}$, consisting of all constant functions, $f_c(x) = c$ for all x. We can use any coding scheme and for any tampering

function $f_c \in \mathcal{F}_{constant}$, we may take D_{f_c} to be $\mathrm{Dec}(c)$ with probability 1.

Note that there cannot exist a code with block length n which is non-malleable with respect to $\mathcal{F}_n$ (recall this is family of all functions from n bits to n bits). This follows since the tampering function could then use the function Dec to decode the message m, get a message m' by flipping all the bits in m, and use the encoding function to pick any codeword in $\mathrm{Enc}(m')$.

Therefore, it is natural to restrict the size of the family of tampering functions. It follows from the works in [DPW10, CG14a] that there exists non-malleable codes with respect to any tampering function family of size bounded by $2^{2^{\delta n}}$ with rate close to $1-\delta$ and error $2^{-\Omega(n)}$, for any constant $\delta > 0$. The bounds obtained in these works are existential, and some progress has been made since then in giving explicit constructions against useful classes of tampering functions.

12.0.1 Non-malleable Codes in the Split-State Model

An important and well studied family of tampering functions (which is also relevant to the current work) is the family of tampering functions in the C-split-state model, for $C \geq 2$. In this model, each tampering function f is of the form $(f_1, \ldots, f_C)$ where $f_i \in \mathcal{F}_{n/C}$, and for any codeword $x = (x_1, \ldots, x_C) \in (\{0,1\}^{n/C})^C$ we define $(f_1, \ldots, f_C)(x_1, \ldots, x_C) = (f_1(x_1), \ldots, f_C(x_C))$. Thus each f_i independently tampers a fixed partition of the codeword. Non-malleable codes in this model can also be viewed as *non-malleable secret sharing*. This is because the strings $(x_1, \ldots, x_C)$ can be seen as the shares of s and tampering each share individually does not allow one to "maul" the shared secret s.

There has been a lot of recent work on constructing explicit and efficient non malleable codes in the C-split-state model. Since $C = 1$ includes all of $\mathcal{F}_n$, the best one can hope for is $C = 2$. A Monte-Carlo construction of non-malleable codes in this model was given in the original paper on non-malleable codes [DPW10] for $C = 2$ and then improved in [CG14a]. However, both of these constructions are inefficient. For $C = 2$, these Monte-Carlo constructions imply existence of codes of rate close to $\frac{1}{2}$ and corresponds to the hardest case. On the other extreme, when $C = n$, it corresponds to the case of bit tampering where each function f_i acts independently on a particular

bit of the codeword.

The best known explicit construction of non-malleable codes in the C-split-state model for the case when $C = 2$ is due to the work of Aggarwal, Dodis and Lovett [ADL14], who construct a code with rate $= \Omega(n^{-6/7})$ and error $= 2^{-\Omega(n^{-1/7})}$. Their proof of non-malleability uses methods from additive combinatorics. The drawback of this construction is the polynomially small rate of the code.

12.0.2 Our Result

The results in this chapter are based on joint works with Vipul Goyal, Xin Li, and David Zuckerman [CZ14, CGL16].

Our main result on non-malleable codes is for the model of C-split-state adversaries when $C = 10$. We give explicit constructions of non-malleable codes in this model with rate $= \Omega(1)$ and error $= 2^{-\Omega(n)}$. In particular, we have the following result.

Theorem 29. *For all $n > 0$ there exists an explicit construction of efficient non-malleable codes on $\{0,1\}^n$ in the 10-split-state model with constant rate and error $= 2^{-\Omega(n)}$.*

We note that the best known non-malleable code in the $O(1)$-split-state prior to this work was the non-malleable code in the 2-split-state model from [ADL14], which as mentioned above, has rate $\Omega(n^{-6/7})$ and error is $2^{-\Omega(n^{-1/7})}$. Thus we give the first explicit construction of constant rate non-malleable codes in the split-state model for a fixed integer C that do not rely on any unproven assumptions; in fact, this is the first for $C = o(n)$. We further obtain optimal error.

For the case of bit tampering ($C = n$), the best known explicit constructions of non-malleable codes were given in the work of [CG14b] with rate $= (1 - o(1))$ and error $= 2^{-\Omega(n^{-1/7})}$. We improve upon the error and obtain the following result.

Theorem 30. *For all $n > 0$ there exists an explicit construction of efficient non-malleable codes on $\{0,1\}^n$ in the bit tampering model with rate $= (1 - o(1))$ and error $= 2^{-\Omega(n)}$.*

We obtain Theorem 30 from the following observation. The construction against bit tampering in [CG14b] uses a possibly sub-optimal rate non-malleable code against bit-tampering in its construction and shows a way to improve the rate to $(1 - o(1))$ while maintaining the error bound. The sub-optimal rate non-malleable code used was the code from [ADL14] which resulted in the sub-optimal error bound of $2^{-\Omega(n^{-1/7})}$. By plugging in our non-malleable code construction from Theorem 29 as the sub-optimal non-malleable code in the construction of [CG14b], we deduce Theorem 30.

Subsequent Work: Aggarwal et al. [ADKO15] constructed explicit non-malleable codes in the 2-split model with constant rate and optimal error. A crucial part of their contruction is our 10-split-state non-malleable code from Theorem 29.

Other Results on Non-Malleable Codes Apart from the previous work stated above, there has been other work in constructing non-malleable codes. However they did not improve the parameters achieved in [ADL14] in the C-split model for $C = o(n)$. Before the work of [ADL14], the only unconditional efficient non-malleable code in the C-split-state model, for $C = o(n)$, was by Dziembowski, Kazana, and Obremski [DKO13]. However, they could encode only 1 bit messages.

In the model of global tampering, Agrawal et al. [AGM$^+$14] constructed efficient non-malleable codes with rate $1 - o(1)$ against the family of permutations. There were also some conditional results. Liu and Lysyanskaya [LL12] constructed efficient constant rate non-malleable codes in the split-state model against computationally bounded adversaries. Their proof of non-malleability relies on the existence of robust public-key cryptosystems and existence of robust non-interactive zero-knowledge proof systems for some language in NP. They also use the common reference string (CRS) assumption which roughly states that one has access to an untampered random string. The work of Faust et al. [FMVW13] constructed almost optimal non-malleable codes against the class of polynomial sized circuits in the CRS framework. [CCP12, CCFP11, CKM11, FMNV14] considered non-malleable codes in other models.

12.1 Multi-Tampered Non-Malleable Codes

We introduce the notion of non-malleable codes that can handle multiple tamperings in the information theoretic setting.

Definition 12.1.1 (One-Many Non-malleable codes)**.** *A coding scheme* $(\mathrm{Enc}, \mathrm{Dec})$ *with block length* n *and message length* k *is a non-malleable code with respect to a family of tampering functions* $\mathcal{F} \subset (\mathcal{F}_n)^t$ *and error* ϵ *if for every* $(f_1, \ldots f_t) \in \mathcal{F}$*, there exists a random variable* $D_{\vec{f}}$ *on* $(\{0,1\}^k \cup \{same^\star\})^t$ *which is independent of the randomness in* Enc *such that for all messages* $s \in \{0,1\}^k$*, it holds that*

$$|(\mathrm{Dec}(f_1(\mathbf{X})), \ldots, \mathrm{Dec}(f_t(\mathbf{X}))) - \mathrm{copy}^{(t)}(D_{\vec{f}}, s)| \leq \epsilon$$

where $\mathbf{X} = \mathrm{Enc}(s)$*. We refer to* t *as the tampering degree of the non-malleable code.*

Thus one-many non-malleable codes is a natural more robust version of the well studied notion of non-malleable codes, and can be used in all applications of non-malleable codes in tamper-resilient cryptography with this stronger form of security.

An expert in cryptography by now would have noticed this is analogous to the well studied notion of one-many non-malleable *commitments* [PR08]. Even though both notions deal with related concerns, we note non-malleable codes and non-malleable commitment are fundamentally different objects with the latter necessarily based on complexity assumptions. To start with, we prove a simple impossibility result for one-many non-malleable codes (whereas for one-many non-malleable commitments, a corresponding *positive* result is known [PR08]).

Lemma 12.1.2. *One-many non-malleable codes which work for any arbitrary tampering degree and* $\epsilon < 1/4$ *cannot exist for a large class of tampering functions.*

Proof. The class of tampering functions which we consider are the ones where each function is allowed to read any one bit $\mathbf{X}_i$ of its choice from the input code $\mathbf{X}$, and output a fresh encoding of $\mathbf{X}_i$. Most natural tampering functions (including split state ones [DPW10, CG14a]) considered

in the literature fall into this class. Assume that the encoded value s has at least 4 possibilities (length 2 bits or higher). The case of a single bit s is discussed later.

Recall that n is the length of the code. We set $t = n$. Let $\mathbf{X} = \text{Enc}(s)$ be the input codeword where s is chosen at random. We consider n tampering functions where F_i simply reads $\mathbf{X}_i$ and outputs a fresh encoding $\mathbf{W}_i = \text{Enc}(\mathbf{X}_i)$. Now consider $(\text{Dec}(f_1(\mathbf{X})), \ldots, \text{Dec}(f_n(\mathbf{X})))$. Observe that this is exactly the bits of the string $\mathbf{X}$. If the distinguisher applies the decode procedure on $\mathbf{X}$, it will recover s. Now consider any possible output $(d_1, \ldots, d_n)$ of $D_{\vec{f}}$. Now note that there cannot exist d_i which is $same^\star$. This is because otherwise it will be replaced by s (see Definition 12.1.1) which is at least 2 bits while $\text{Dec}(W_i)$ is just a single bit. This in turn implies that the value $\text{copy}(D_{\vec{f}}, s)$ (from Definition 12.1.1) is independent of s and $\mathbf{X}$. Thus a distinguisher (given access to s) can easily have an advantage exceeding ϵ.

For a single bit s, we modify our tampering functions to encode two bits: $\mathbf{W}_i = \text{Enc}(\mathbf{X}_i||0)$. Then again we can argue that neither of d_i will be $same^\star$ since then it will be replaced by s which is only one bit. This in turn again implies that $\text{copy}(D_{\vec{f}}, s)$ is independent of s and X. This concludes the proof.

□

We also introduce a natural generalization which we call *many-many non-malleable codes*. This refers to the situation where the adversary is given multiple codewords as input.

Definition 12.1.3 (Many-Many Non-malleable codes)**.** *A coding scheme* (Enc, Dec) *with block length* n *and message length* k *is a non-malleable code with respect to a family of tampering functions* $\mathcal{F} \subset (\mathcal{F}_n)^t$ *and error* ϵ *if for every* $(f_1, \ldots f_t) \in \mathcal{F}$*, there exists a random variable* $D_{\vec{f}}$ *on* $(\{0,1\}^k \cup \{same^\star{}_i\}_{i\in[u]})^t$ *which is independent of the randomness in* Enc *such that for all vector of messages* $(s_1, \ldots, s_u)$*,* $s_i \in \{0,1\}^k$*, it holds that*

$$|(\text{Dec}(f_1(\vec{\mathbf{X}})), \ldots, \text{Dec}(f_t(\vec{\mathbf{X}}))) - \text{copy}(D_{\vec{f}}, (s_1, \ldots, s_u))| \leq \epsilon$$

Where $X_i = \text{Enc}(s_i)$ *and* $\vec{X} = (X_1, \ldots, X_u)$

The following lemma relates one-many non-malleable codes to many-many non-malleable codes. This lemma is analogous to a similar lemma for non-malleable commitments [PR08].

Lemma 12.1.4. *One-many non-malleable codes with tampering degree t and error ϵ are also many-many non-malleable codes for tampering degree t and error $u\epsilon$ (where u is as in Definition 12.1.3).*

Proof. This proof relies on a simple hybrid argument and the fact that all sources $\mathbf{X}_1, \ldots, \mathbf{X}_u$ are independent. We only provide a proof sketch here. Assume towards contradiction that there exists a one-many code with error ϵ, which, under the many-many tampering adversary has error higher than $u.\epsilon$. That is, the adversary $\vec{(f)}$ is given as input $(\mathbf{X}_1, \ldots, \mathbf{X}_u)$ which are encodings of $(s_1, \ldots, s_u)$ respectively. This is referred to as the hybrid 0. Now consider the following hybrid experiment. In the i-th hybrid experiment, the code X_i is changed to be an encoding of 0 (as opposed to be an encoding of s_i). We claim that in this experiment, the error changes by at most ϵ. This is because otherwise we can construct a one-many tampering adversary with error higher than ϵ. To construct such an adversary $\vec{(f^i)}$, each f^i_j has $X_{kk\neq i}$ hardcoded in it and takes X_i as input. This would show an adversary against which one-many non-malleable codes have an error higher than ϵ.

By the time we reach $(u-1)$-th hybrid experiment, the error could only have reduced by at most $(u-1)\epsilon$. However in the $(u-1)$-th hybrid experiment, the error can at most be ϵ since it corresponds to the one-many setting. Hence, the error in the hybrid 0 could have been at most $u.\epsilon$. This concludes the proof.

□

Our main result is the following.

Theorem 31. *There exists a constant $\gamma > 0$ such that for all $n > 0$ and $t \leq n^\gamma$, there exists an efficient construction of one-many non-malleable codes in the 2-split state model with tampering degree t, relative rate $n^{\Omega(1)}/n$, and error $2^{-n^{\Omega(1)}}$.*

Relation to Continuous Non-Malleable Codes A primitive related to one-many non-malleable codes that we introduce, known as continuous non-malleable codes, was introduced by Faust et al. [FMNV14]. Informally, in a continuous non-malleable code, the codewords are allowed to be tampered multiple times (without allowing fresh encoding of the message), with the additional guarantee that the tampering experiment stops (called "self destruct") whenever an error message is detected. This model is weaker than the notion we consider since we do not allow for such a self-destruct option. However the work of [FMNV14] allows for unbounded number of tamperings. On the other hand, their constructions are based on computational assumptions while ours are purely information-theoretic.

The work of Jafargholi and Wichs [JW15] studied variants of continuous non-malleable codes, depending on whether the tampering is persistent (i.e., the new tampering is on the current tampered version of the codeword) or non-persistent (i.e., the tampering is always on the original codeword). Further [JW15] considered variants depending on whether the self-destruct option is available.

It was shown in [FMNV14] that continuous non-malleable codes against unbounded tampering in the non-persistent model cannot exist in the information theoretic setting. Subsequently, the work of [JW15] proved the existence of continuous non-malleable codes against unbounded tampering in the persistent model (with self-destruct) in the information theoretic setting. Following this, in a recent work Aggarwal, Kazana and Ombreski [AKO15] provided explicit constructions of such codes.

Thus, our result on one-many non-malleable codes can be interpreted as an explicit construction of continuous non-malleable codes in the non-persistent model (without self-destruct) against a bounded tampering in the information-theoretic model. We note that as implied by the result of [FMNV14], one cannot hope to handle unbounded tampering in this model in the information theoretic setting.

12.2 Non-malleable codes via Seedless non-malleable extractors

Seedless non-malleable extractors were introduced by Cheraghchi and Guruswami in [CG14b], where it was shown that explicit constructions of such extractors can be used to construct non-malleable codes[2].

The following theorem generalizes a result of Cheraghchi and Gursuswami [CG14b].

Theorem 12.2.1. *Let* nmExt : $\{0,1\}^n \times \{0,1\}^n \to \{0,1\}^m$ *be a polynomial time computable seedless* $(2,t)$*-non-malleable extractor for independent sources at min-entropy* $n - n^\gamma$ *with error* ϵ*. Then there exists an explicit non-malleable code with an efficient decoder in the* $(2,t)$*-split-state model with block length* $= 2n$*, rate* $= \frac{m}{2n}$ *and error* $= 2^{(m+2)t}(\epsilon + 2^{-n^\gamma})$*.*

Proof. Let $\mathcal{A}_1 = (f_1, g_1), \ldots, \mathcal{A}_t = (f_t, g_t)$ be arbitrary 2-split-state adversaries. We partition $\{0,1\}^n$ in two different ways based on the fixed points of the tampering functions.

For any $R \subseteq [t]$, define

$$W^{(R)} = \{x \in \{0,1\}^n : f_i(x) = x \text{ if } i \in R, \text{ and } f_i(x) \neq x \text{ if } i \in [t] \setminus R\}.$$

Similalry, for any $S \subseteq [t]$, define

$$V^{(S)} = \{y \in \{0,1\}^n : g_i(y) = y \text{ if } i \in S, \text{ and } g_i(y) \neq y \text{ if } i \in [t] \setminus S\}.$$

Thus the sets $W^{(R)}, R \subseteq [t]$ defines a partition of $\{0,1\}^n$. Similarly $\mathbf{V}^{(S)}, S \subseteq [t]$ defines a partition of $\{0,1\}^n$. For $R, S \subseteq [t]$, let $\mathbf{X}^{(R)}$ be a random variable uniform on $\mathbf{W}^{(R)}$, and $\mathbf{Y}^{(S)}$ be a random variable uniform on $\mathbf{V}^{(S)}$.

Let $\mathbf{U}_{n_4}$ be uniform on $\{0,1\}^{n_4}$ and independent of $\mathbf{X}^R, \mathbf{Y}^S$, for all $R, S \subseteq [t]$.

Define

$$D^{(R,S)}_{\vec{f},\vec{g}} = (\mathbf{U}_{n_4}, Z_1^{(R,S)}, \ldots, Z_t^{(R,S)})$$

[2]the encoder of the resulting non-malleable code may still be inefficient. Informally, to make the encoder efficient, one needs to sample from the pre-image of any output of the extractor. See Section 12.4 for more details.

where we define the random variable

$$Z_i^{(R,S)} = \begin{cases} \text{nmExt}(f_i(X^{(R)}), g_i(Y^{(S)})) & \text{if } i \in [t] \setminus (R \cap S) \\ same^\star & \text{if } i \in R \cap S \end{cases}$$

Define the distribution:

$$D_{\vec{f},\vec{g}} = \sum_{R,S} \alpha_{R,S} D_{\vec{f},\vec{g}}^{(R,S)}$$

, where $\alpha_{R,S} = \frac{|W^{(R,S)}||V^{(R,S)}|}{2^{2n}}$.

We first prove the following claim.

Claim 12.2.2. *Let*

$$\Delta_{R,S} = \alpha_{R,S} |\text{nmExt}(\mathbf{X}^{(R)}, \mathbf{Y}^{(S)}), \text{nmExt}(f_1(\mathbf{X}^{(R)}), g_1(\mathbf{Y}^{(S)})), \ldots, \text{nmExt}(f_t(\mathbf{X}^{(R)}), g_t(\mathbf{Y}^{(S)})) - D_{\vec{f},\vec{g}}^{(R,S)}|.$$

Then, for every $R, S \subseteq [t]$, $\Delta_{R,S} \leq 2^{-n^\gamma} + \epsilon$.

Proof. If $|\mathbf{W}^{(R)}| \leq 2^{n-n^\gamma}$, it follows that $\alpha_{R,S} \leq 2^{-n^\gamma}$, and hence the claim follows. Thus, assume that $H_\infty(\mathbf{X}^{(R)}) \geq n - n^\gamma$. Using a similar argument, we can assume that $H_\infty(\mathbf{Y}^{(S)}) \geq n - n^\gamma$.

Let $\overline{R \cap S} = [t] \setminus (R \cap S) = \{i_1, \ldots, i_j\}$. It follows that for any $c \in \overline{R \cap S}$, at least one the following is true: (1) f_c has no fixed points on $W^{(R)}$ (2) g_c has no fixed points on $V^{(S)}$. Thus, using the fact nmExt $(2,t)$-non-malleable extractor, we have

$$|\text{nmExt}(\mathbf{X}^{(R)}, \mathbf{Y}^{(S)}), \text{nmExt}(f_{i_1}(\mathbf{X}^{(R)}), g_{i_1}(\mathbf{Y}^{(S)})), \ldots, \text{nmExt}(f_{i_j}(\mathbf{X}^{(R)}), g_{i_j}(\mathbf{Y}^{(S)})) - \mathbf{U}_{n_4}, \text{nmExt}(f_{i_1}(\mathbf{X}^{(R)}), g_{i_1}(\mathbf{Y}^{(S)})), \ldots, \text{nmExt}(f_{i_j}(\mathbf{X}^{(R)}), g_{i_j}(\mathbf{Y}^{(S)}))| \leq \epsilon$$

The claim now follows by observing that for each $c \in R \cap S$, f_c and g_c are the identity functions on the sets $\mathbf{W}^{(R)}$ and $\mathbf{V}^{(S)}$ respectively. □

Let $\mathbf{X}, \mathbf{Y}$ be independent and uniformly random on $\{0,1\}^n$. Thus, we have

$$|\mathrm{nmExt}(\mathbf{X},\mathbf{Y}), \mathrm{nmExt}(\mathcal{A}_1(\mathbf{X},\mathbf{Y})), \ldots, \mathrm{nmExt}(\mathcal{A}_t(\mathbf{X},\mathbf{Y}))$$
$$-\mathbf{U}_{n_4}, \mathrm{copy}^{(t)}(D_{\vec{f},\vec{g}}, \mathbf{U}_{n_4})| = \sum_{R,S\subseteq[t]} \Delta_{R,S} \leq 2^{2t}(\epsilon + 2^{-n^\gamma}).$$

We now define the non-malleable code in the following way: For any message $s \in \{0,1\}^m$, the encoder $\mathrm{Enc}(s)$ outputs a uniformly random string from the set $\mathrm{nmExt}^{-1}(s) \subset \{0,1\}^{Cn}$. For any codeword $c \in \{0,1\}^{Cn}$, the decoder Dec outputs $\mathrm{nmExt}(c)$. It follows that for any t-tuple of messages : $(s_1, \ldots, s_t) \in (\{0,1\}^m)^t$, we have

$$\begin{aligned}|(\mathrm{Dec}(f_1(\vec{\mathbf{X}})), \ldots, \mathrm{Dec}(f_t(\vec{\mathbf{X}}))) - \mathrm{copy}(D_{\vec{f},\vec{g}}, (s_1, \ldots, s_t))| &\leq 2^{mt}|\mathrm{nmExt}(\mathbf{X},\mathbf{Y}), \mathrm{nmExt}(\mathcal{A}_1(\mathbf{X},\mathbf{Y})),\\ &\quad \ldots, \mathrm{nmExt}(\mathcal{A}_t(\mathbf{X},\mathbf{Y})) - \mathbf{U}_{n_4}, \mathrm{copy}^{(t)}(D_{\vec{f},\vec{g}}, \mathbf{U}_{n_4})|\\ &\leq 2^{mt+2t}(\epsilon + 2^{-n^\gamma}).\end{aligned}$$

□

Remark 12.2.3. *Thus, note that to construct efficient non-malleable codes using a seedless non-malleable extractor* nmExt*, we also need to sample efficiently from a distribution that is almost uniform on* $\mathrm{nmExt}^{-1}(s)$ *for any message* s*.*

12.3 Efficient algorithms for non-malleable codes in the 10-split-state model

In this section we prove efficiency of the non-malleable codes in the 10-split-state model that follow via the non-malleable extractor construction in Section 11.2 (using Theorem 12.2.1). Recall that for any message s, its encoding is a uniform element from $\mathrm{nmExt}^{-1}(s)$ and for any codeword c, the decoded message is $\mathrm{nmExt}(c)$. Thus the efficiency of the decoder follows from the fact that nmExt is a polynomial time function.

We construct an efficient algorithm which takes as input a message $s \in \{0,1\}^n$ and samples from a distribution that is $2^{-\Omega(n)}$-close to uniform on $\mathrm{nmExt}^{-1}(s)$ and use this as our encoder. This is indeed sufficient, since we only add an exponentially small error when we use this algorithm instead of sampling uniformly from $\mathrm{nmExt}^{-1}(s)$.

Our sampling algorithm is based on the following observations.

- The uniform distribution on the set $\mathrm{nmExt}^{-1}(s)$ is a convex combination of uniform distributions on algebraic varieties of low degree.
- Sampling almost uniformly from such algebraic sets can be done efficiently [CS09].
- Further, obtaining the weights in the convex combination reduces to approximately counting the size of such algebraic sets for which there are efficient algorithms [HW98]. However, the number of distributions in the convex combination can be exponentially large. To get around this difficulty, we use the method of rejection sampling. The proof of correctness of the algorithm relies on estimates on the number of rational points on algebraic varieties.

12.3.1 Tools from algebraic geometry

Let $g \in \mathbb{F}_p[x_1, \ldots, x_c]$ and let $\mathcal{H} \subseteq \mathbb{F}_p^c$ be its set of zeroes. We call $\mathcal{H}$ the algebraic hypersurface defined by g.

The following version of the Lang-Weil bound for hypersurfaces in $\mathbb{F}_p^c$ was proved by Cafure and Matera [CM06].

Theorem 12.3.1 (Lang-Weil bound)**.** *Let c, d be constant integers and let p be a large prime. Let $\mathcal{H} \subset \mathbb{F}_p^c$ be a hypersurface defined by a degree d polynomial. Then there exists an integer s, $0 \leq s \leq d$, such that*

$$||\mathcal{H}| - sp^{c-1}| \leq O(\mathrm{sign}(s) \cdot p^{c-\frac{3}{2}} + p^{c-2})$$

where $\mathrm{sign}(s) = 1$ *if* $s > 0$ *and* $\mathrm{sign}(0) = 0$.

Lemma 12.3.2 (Schwartz-Zippel Lemma [Sch80, Zip79]). *Let $g(x_1, \ldots, x_c)$ be a non-zero multivariate polynomial of degree d with coefficients in $\mathbb{F}_p$. Then the hypersurface $\mathcal{H} \subset \mathbb{F}_p^c$ defined by g is of size at most dp^{c-1}.*

We need some previous work on efficient sampling and approximate counting of algebraic varieties.

Theorem 12.3.3 ([CS09]). *Let c, k, d be constant integers such that $c > k$ and let p be a prime. There exists an efficient randomized algorithm $\mathcal{A}_1$ such that the following holds:*

Let $g_1, \ldots, g_k \in \mathbb{F}_p[x_1, \ldots, x_c]$ be arbitrary polynomials of degree at most d and let $S \subseteq \mathbb{F}_p^c$ be the set of common zeroes of $g_1, \ldots, g_k$. $\mathcal{A}_1$ takes as input the description of $g_1, \ldots, g_k$ and a parameter δ and outputs a sample from a distribution which is $O(1/p^{1-\delta})$-close to the uniform distribution on S. The worst-case running time of $\mathcal{A}_1$ is bounded by $\mathrm{poly}(\log p)$.

Theorem 12.3.4 ([HW98]). *Let $c, k, d > 0$ be constant integers and let p be a prime. There exists an efficient randomized algorithm $\mathcal{A}_2$ such that the following holds:*

Let $g_1, \ldots, g_k \in \mathbb{F}_p[x_1, \ldots, x_c]$ be arbitrary polynomials of degree at most d and let $S \subseteq \mathbb{F}_p^c$ be the set of common zeroes of $g_1, \ldots, g_k$. $\mathcal{A}_2$ takes as input the description of $g_1, \ldots, g_k$ and outputs an integer v such that

$$\frac{1}{|S|} \cdot |v - |S|| < O(p^{-1/2})$$

The worst-case running time of $\mathcal{A}_2$ is bounded by $\mathrm{poly}(\log p)$.

12.3.2 A new extractor

In the construction of the seedless non-malleable extractor nmExt in Section 11.2, we needed a seeded non-malleable extractor snmExt (with some additional properties, see Theorem 11.2.8). We carefully choose snmExt such that it is easy to sample almost uniformly from $\mathrm{nmExt}^{-1}(s)$. The main idea is to pick snmExt such that $\mathrm{nmExt}^{-1}(s)$ is a convex combination of algebraic varieties of low degree over a field with large characteristic. Thus, the constructions in [Li12b] look to be a

good choice for the seeded non-malleable extractor. However, for this choice, we face the following difficulty:

Let $\sigma_M : \mathbb{F}_p \to \mathbb{Z}_M$ be defined as $\sigma_M(x) = x \pmod M$. nmExt is of the form $\sigma_M \circ \text{ext}_2 \circ \text{ext}_1$, where $\text{ext}_1 : \mathbb{F}_p^{10} \to \mathbb{F}_p^4$, $\text{ext}_2 : \mathbb{F}_q^2 \to \mathbb{F}_q$, and p, q are primes satisfying $p^2 \leq q \leq 2p^2$ (and interpreting the output of ext_1 as an element in $\mathbb{F}_q^2$). Changing the characteristic of the field destroys the low degree properties of the function $\text{ext}_2 \circ \text{ext}_1$.

To fix this, we construct a new extractor for ext_2 (satisfying the conditions of Theorem 11.2.8) which allows us to work over the same field as ext_1. The extractor is a variation of a construction by Bourgain [Bou05b]. The proof is easy to obtain by using ideas from [Bou05b,Li12b], and we omit it.

Theorem 12.3.5. *Let p be a prime. Define the functions $\text{ext}_2 : (\mathbb{F}_p^2) \times (\mathbb{F}_p^2) \to \mathbb{F}_p$ and $\text{snmExt} : (\mathbb{F}_p^2) \times (\mathbb{F}_p^2) \to \mathbb{Z}_M$ in the following way:*

$$\text{ext}_2((x_1, x_2), (y_1, y_2)) = \sum_{j=1}^{2} (x_j y_j + x_j^2 y_j^2), \qquad \text{snmExt}(x, y) = \sigma_M(\text{ext}_2(x, y))$$

where $\sigma_M(x) = x \pmod M$. Suppose $\mathbf{X}, \mathbf{Y}$ are independent sources on $\mathbb{F}_p^2$ with min-entropies k_1, k_2 respectively.

1. *If $(k_1 + k_2) \geq (2 + \delta) \log p$, then*

$$|\text{snmExt}(\mathbf{X}, \mathbf{Y}) \circ \mathbf{X} - U_M \circ \mathbf{X}| < p^{-\Omega(1)}, \ |\text{snmExt}(\mathbf{X}, \mathbf{Y}) \circ \mathbf{Y} - U_M \circ \mathbf{Y}| < p^{-\Omega(1)}$$

2. *If $k_1, k_2 > (2 - \delta) \log p$ and f is any tampering function with no fixed points, then*

$$|\text{snmExt}(\mathbf{X}, \mathbf{Y}) \circ \text{snmExt}(\mathbf{X}, f(\mathbf{Y})) - U_M \circ \text{snmExt}(\mathbf{X}, f(\mathbf{Y}))| < p^{-\Omega(1)}.$$

12.3.3 A generic sampling algorithm

We construct an algorithm for almost uniformly sampling from certain structured sets.

Theorem 12.3.6. *Let S_1, S_2, and S_3 be finite sets. For arbitrary functions $g : S_2 \to S_3$, $h : S_1 \to S_2$, there exists a sampling algorithm $\mathcal{B}$ which takes as input $z \in S_3$ and a parameter $\epsilon \geq \epsilon_0$, runs in time* $\mathrm{poly}(\log(|S_1| \cdot |S_2|), \log(\frac{1}{\epsilon}))$*, and outputs a sample from a distribution that is $O(\epsilon)$-close to uniform on the set $(g \circ h)^{-1}(z)$, if the following conditions hold:*

1. *There exists an algorithm $\mathcal{B}_1$, which takes as input $z \in S_3$, runs in time* $\mathrm{poly}(\log(|S_2|))$*, and outputs a sample from a distribution that is uniform on the set $g^{-1}(z)$.*

2. *There exists an algorithm $\mathcal{B}_2$, which takes as input $y \in S_2$ and ϵ, runs in time* $\mathrm{poly}(\log(|S_1|), \log(\frac{1}{\epsilon}))$*, and outputs a sample from a distribution that is ϵ-close to uniform on the set $h^{-1}(y)$.*

3. *There exists an algorithm $\mathcal{B}_3$, which takes as input $y \in S_2$ and ϵ, runs in time* $\mathrm{poly}(\log(|S_1|), \log(\frac{1}{\epsilon}))$*, and outputs an approximation A_y for $|h^{-1}(y)|$ with a multiplicative error of at most ϵ, i.e., $1 - \epsilon \leq \frac{A_y}{|h^{-1}(y)|} \leq 1 + \epsilon$.*

4. *There exist constants $\beta > 0$ and $\lambda \geq 1$, and an efficiently computable value* max *such that for all $\epsilon \geq \epsilon_0$ the following holds: There exists a subset $S_2' \subseteq S_2$ such that for all $y \in S_2'$, $\frac{\max}{\lambda} \leq |h^{-1}(y)| \leq \max$. Further, $\frac{1}{|(g \circ h)^{-1}(z)|} \sum_{y \in S_2 \setminus S_2'} |h^{-1}(y)| \leq \epsilon$ and $\frac{|S_2'|}{|S_2|} > \beta$.*

Proof. The idea is to use the method of rejection sampling.

Algorithm $\mathcal{B}$ (given input $z \in S_3$ and error parameter ϵ):

1. Use $\mathcal{B}_1$ to sample y from $g^{-1}(z)$. Compute an approximation A_y for $|h^{-1}(y)|$ with error ϵ using algorithm $\mathcal{B}_3$. If $A_y < \max \cdot (\frac{1}{\lambda} - \epsilon)$, reject y. Else accept y with probability $wt(y) = \frac{A_y}{\max}$.

   Iterate this step till some y is accepted. If no sample is accepted after $O(\log \frac{1}{\epsilon})$ iterations, accept the next sample.

2. Once y is accepted, sample from $h^{-1}(y)$ using $\mathcal{B}_2$ (with error ϵ).

Proof of correctness of Algorithm $\mathcal{B}$: Consider any subset $T \subseteq (g \circ h)^{-1}(z)$. Let $p_{T,1}$ be the probability that some element from T is picked by $\mathcal{B}$ in one iteration.

Then:

$$p_{T,1} = \sum_{y \in g^{-1}(z)} \frac{1}{|g^{-1}(z)|} \cdot \left(\frac{|h^{-1}(y)|}{\max} \pm \epsilon \right) \cdot \left(\frac{|T \cap h^{-1}(y)|}{|h^{-1}(y)|} \pm \epsilon \right)$$

The above expression is derived in the following way: Consider any $y \in g^{-1}(z)$. Let A_y be the approximation of $|h^{-1}(y)|$ computed by algorithm $\mathcal{B}_3$. The probability of y being picked by $\mathcal{B}_1$ is $\frac{1}{|g^{-1}(z)|}$. The probability that this y is accepted is given by $\frac{A_y}{\max} = \frac{|h^{-1}(y)|}{\max} \pm \epsilon$. Further, if y is accepted, $\frac{|T \cap h^{-1}(y)|}{|h^{-1}(y)|} \pm \epsilon$ is the probability that some element from the set T is picked by algorithm $\mathcal{B}_2$ (since $\mathcal{B}_2$ samples from a distribution ϵ-close to uniform on $h^{-1}(y)$).

It follows that,

$$|p_{T,1} - \frac{|T|}{\max \cdot |g^{-1}(z)|}| = O(\epsilon)$$

Let $N = |(g \circ h)^{-1}(z)|$. The probability that an iteration of Step (1) fails to accept a sample is:

$$p_{reject} = \left(1 - \frac{N}{\max \cdot |g^{-1}(z)|} \right) \pm O(\epsilon)$$

Let $k = O(\log \frac{1}{\epsilon})$. The probability p_T that some element from T is picked by $\mathcal{B}$ in at most k iterations is given by:

$$\begin{aligned} p_T &= p_{T,1} \sum_{i=0}^{k-1} (p_{reject})^i \\ &= \left(\frac{|T|}{\max \cdot |g^{-1}(z)|} \pm O(\epsilon) \right) \cdot \sum_{i=0}^{k-1} \left(1 - \frac{N}{\max \cdot |g^{-1}(z)|} \pm O(\epsilon) \right)^i \end{aligned}$$

Thus,

$$\begin{aligned} \left| p_T - \frac{|T|}{N} \right| &\leq \left(1 - \frac{N}{\max \cdot |g^{-1}(z)|} \right)^k + O(\epsilon) \\ &\leq e^{-\frac{Nk}{\max \cdot |g^{-1}(z)|}} + O(\epsilon) = O(\epsilon) \end{aligned}$$

where the equality in the last step follows from the fact that $\frac{N}{\max \cdot |g^{-1}(z)|} = O(1)$ (by Condition (4) in the hypothesis).

The probability that no sample is accepted by $\mathcal{B}$ in k iterations is bounded by:

$$\left(1 - \frac{N}{\max \cdot |g^{-1}(z)|}\right)^k + O(\epsilon) = O(\epsilon)$$

Let $\mathcal{B}(z, \epsilon)$ denote the output distribution of algorithm $\mathcal{B}$. Thus,

$$\begin{aligned} |\mathcal{B}(z,\epsilon) - U_{(g\circ h)^{-1}(z)}| &= \max_{T \subseteq (g\circ h)^{-1}(z)} \left| \Pr[\mathcal{B}(z,\epsilon) \in T] - \frac{|T|}{N} \right| \\ &\leq \left| p_T + O(\epsilon) - \frac{|T|}{N} \right| = O(\epsilon) \end{aligned}$$

□

12.3.4 An efficient encoder

We recall the seedless non-malleable extractor constructed in Theorem 3.

Let $\text{enc} : \mathbb{F}_p \to \mathbb{F}_p^2$ be defined as $\text{enc}(x) = (x, x^4 + x^2 + x)$.

Then $\text{nmExt} : \mathbb{F}_p^{10} \to \mathbb{Z}_M$ is defined to be:

$$\text{nmExt}(x_1, \ldots, x_{10}) = \text{ext}_3(\text{ext}_2(\text{ext}_1(x_1, \ldots, x_{10})))$$

where, $\text{ext}_1 : \mathbb{F}_p^{10} \to \mathbb{F}_p^4$, $\text{ext}_2 : \mathbb{F}_p^4 \to \mathbb{F}_p$, and $\text{ext}_3 : \mathbb{F}_p \to \mathbb{Z}_M$ are defined in the following way:

$$\text{ext}_1(x_1, \ldots, x_{10}) = \big(\sum_{i=0}^{1} (\text{enc}(x_{4i+1}) + \text{enc}(x_{4i+2})) \odot (\text{enc}(x_{4i+3}) + \text{enc}(x_{4i+4})), x_9, x_{10} \big)$$

$$\text{ext}_2(y_1, y_2, z_1, z_2) = \sum_{j=1}^{2} (y_j z_j + y_j^2 z_j^2), \;\; \text{ext}_3(w) = \sigma_M(w) = w \pmod M$$

We set $M = p^\delta$ such that the error in the extractor nmExt is $\epsilon = p^{-2\delta}$. Note that, as discussed

before, we use the extractor from Subsection 12.3.2 for ext_2 in nmExt instead of the constructions in [DLWZ14, Li12b].

An efficient encoder for the constructed non-malleable codes in the 10-split-state model follows from the following theorem.

Theorem 12.3.7. *There exists a randomized algorithm which takes as input $z \in \mathbb{Z}_M$ and a parameter $\epsilon > O(p^{-1/2})$ and samples from a distribution $O(\epsilon)$-close to uniform on the set $(\text{nmExt})^{-1}(z)$. The worst case running time of the algorithm is bounded by $\text{poly}(\log p, \log(\frac{1}{\epsilon}))$.*

We prove Theorem 12.3.7 using the following lemma.

Lemma 12.3.8. *For $s \in \mathbb{Z}_M$, let $T_s = \text{ext}_3^{-1}(s) \subset \mathbb{F}_p$ and $S = \text{nmExt}^{-1}(s)$. For $a \in \mathbb{F}_p$, define $W_a = (\text{ext}_2 \circ \text{ext}_1)^{-1}(a) \subset \mathbb{F}_p^{10}$. Define $I_s = \{a \in T_s : \frac{|W_a|}{p^9} \leq 0.9\}$ and $W = \bigcup_{a \in I_s} W_a$.*

Then

$$\frac{|W|}{|S|} < p^{-(1-\delta)}, \quad \frac{|I_s|}{|T_s|} < \frac{18}{19}$$

Proof of Theorem 12.3.7 assuming Lemma 12.3.8. We show that for $g = \text{ext}_3$ and $h = \text{ext}_2 \circ \text{ext}_1$, all the conditions of Theorem 12.3.6 are satisfied.

1. It is easy to uniformly sample from $g^{-1}(z)$.
2. An efficient algorithm for almost uniformly sampling from $h^{-1}(y)$ follows from Lemma 12.3.3.
3. An efficient algorithm for approximately counting $h^{-1}(y)$ follows from Lemma 12.3.4.
4. Using Lemma 12.3.8, we have that for at least $(1/19)^{th}$ fraction of the y's in $g^{-1}(z)$, $0.9p^9 < |h^{-1}(y)| \leq 18p^9$.

   Define $I = \{y \in g^{-1}(z) : |h^{-1}(y)| \leq 0.9p^9\}$. It follows from Lemma 12.3.8 that:

$$\frac{1}{|(g \circ h)^{-1}(z)|} \sum_{y \in I} |h^{-1}(y)| < p^{-(1-\delta)}$$

Thus by Theorem 12.3.6, there exists an efficient algorithm to sample almost uniformly from the set $(\text{nmExt})^{-1}(z)$. □

Proof of Lemma 12.3.8. We begin by proving some claims.

Claim 12.3.9. *For any $s \in \mathbb{Z}_M$,*

$$p^{10-\delta}(1-p^{-\delta}) < |\text{nmExt}^{-1}(s)| < (p^{10-\delta})(1+p^{-\delta})$$

Proof. Let $\mathbf{X}_1, \ldots, \mathbf{X}_{10}$ be uniform on $\mathbb{F}_p$. Using the fact that nmExt is an extractor for independent sources with error at most $\epsilon = p^{-2\delta}$, we have $|\Pr[\text{nmExt}(\mathbf{X}_1, \ldots, \mathbf{X}_{10}) = s] - \frac{1}{M}| < \epsilon$. The bound on $|\text{nmExt}^{-1}(s)|$ now follows. □

Claim 12.3.10. *For any $a \in \mathbb{F}_p$, let $W_a = (\text{ext}_2 \circ \text{ext}_1)^{-1}(a) \subset \mathbb{F}_p^{10}$. Then there exists a polynomial $g \in \mathbb{F}_p[x_1, \ldots, x_{10}]$ of degree at most 18 with coefficients in $\mathbb{F}_p$ such that W_a is the set of zeroes of g.*

Proof. Define $g(x_1, \ldots, x_{10}) = \text{ext}_2 \circ \text{ext}_1(x_1, \ldots, x_{10}) - a$. □

For $a \in \mathbb{F}_p$, define $N_a = |W_a|$. Note that $|T_s| = p^{1-\delta}$.

Using Claim 12.3.9, we have

$$p^{10-\delta} - p^{10-2\delta} \leq \sum_{a \in T_s} N_a \leq p^{10-\delta} + p^{10-2\delta}$$

It follows from Lemma 12.3.2 and Claim 12.3.10 that for any $a \in \mathbb{F}_p$, $N_a \leq 18p^9$. Further, Theorem 12.3.1 and Claim 12.3.10 imply that if $N_a < 0.9p^9$ for some $a \in \mathbb{F}_p$, then $N_a < Cp^8$ for some constant C.

Thus,

$$p^{10-\delta} - p^{10-2\delta} \leq |I_s| \cdot Cp^8 + (|T_s| - |I_s|) \cdot 18p^9 \tag{12.1}$$

for some constant C.

Since $|I_s| \leq |T_s| = p^{1-\delta}$, $|I_s| \cdot Cp^8 \leq Cp^{9-\delta}$. It follows that,

$$p^{1-\delta} - o(1) \leq 18(|T_S| - |I_s|) \tag{12.2}$$

Rearranging, we have

$$\frac{|I_s|}{|T_s|} \leq \frac{17}{18} + o(1) < \frac{18}{19}$$

Further,

$$\frac{|W|}{|S|} < \frac{1}{|S|} \cdot |I_s| \cdot Cp^8 \leq C \cdot \frac{p^{9-\delta}}{p^{10-\delta} - p^{10-2\delta}} < p^{-(1-\delta)}.$$

□

12.4 Efficient Encoding and Decoding Algorithms for One-Many Non-Malleable Codes

In this section, we construct efficient algorithms for almost uniformly sampling from the pre-image of any output of a modified version of the $(2, t)$-non-malleable extractor constructed in Section 11.3. Combining this with Theorem 12.2.1 and Theorem 11.3.1 gives us efficient constructions of one-many non-malleable codes in the 2-split state model, with tampering degree $t = n^{\Omega(1)}$, relative rate $n^{\Omega(1)}/n$ and error $2^{-n^{\Omega(1)}}$.

A major part of this section is on modifying the components used in the construction of nmExt (Algorithm 10) so that the overall extractor is much simpler to analyze as a function, and this enables us to develop efficient sampling algorithms from the pre-image. We present the modified extractor construction in Section 12.4.2. However, we first need to solve a simpler problem.

12.4.1 A New Linear Seeded Extractor

A crucial sub-problem that we have to solve is almost uniformly sampling from the pre-image of a linear seeded extractor in polynomial time. Towards this, we recall a well known property of linear

seeded extractors.

Lemma 12.4.1 ([Rao09b])**.** *Let* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a linear seeded extractor for min-entropy* k *with error* $\epsilon < \frac{1}{2}$*. Let* $\mathbf{X}$ *be an affine* (n,k)*-source. Then*

$$\Pr_{u \sim \mathbf{U}_d}[|\mathrm{Ext}(\mathbf{X}, u) - \mathbf{U}_m| > 0] \leq \epsilon.$$

□

Definition 12.4.2. *For any seeded extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$*, any* $s \in \{0,1\}^d$ *and* $r \in \{0,1\}^m$*, we define:*

- $\mathrm{Ext}(\cdot, s) : \{0,1\}^n \to \{0,1\}^m$ *to be the map* $\mathrm{Ext}(\cdot, s)(x) = \mathrm{Ext}(x, s)$*.*
- $\mathrm{Ext}^{-1}(r)$ *to be the set* $\{(x,y) \in \{0,1\}^n \times \{0,1\}^d : \mathrm{Ext}(x,y) = r\}$*.*
- $\mathrm{Ext}^{-1}(\cdot, s)$ *to be the set* $\{x : \mathrm{Ext}(x,s) = r\}$*.*

We now present a natural way of sampling from pre-images of linear seeded extractors.

Claim 12.4.3. *Let* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a linear seeded extractor for min-entropy* k *with error* $\epsilon < 2^{-1.5m}$*. For any* $r \in \{0,1\}^m$*, consider the following efficient sampling procedure* $\mathcal{S}$ *which on input* r *does the following: (a) Sample* $s \sim \mathbf{U}_d$*, (b) sample* x *uniformly from the subspace* $\mathrm{Ext}(\cdot, s)^{-1}(r)$*. (c) Output* (x, s)*. Let* $\mathcal{D}_r$ *be the distribution uniform on* $\mathrm{Ext}^{-1}(r)$*, and let* $\mathcal{S}(r)$ *denote the distribution produced by* $\mathcal{S}$ *on input* r*.*

Then,

$$|\mathcal{S}(r) - \mathcal{D}_r| \leq 2^{-\Omega(m)}$$

Proof. Define the sets:

$$\mathit{Good} = \{s \in \{0,1\}^d : \mathrm{rank}(\mathrm{Ext}(\cdot, s)) = m\}, \quad \mathit{Bad} = \{0,1\}^d \setminus \mathit{Good}.$$

It follows by Lemma 12.4.1 that $|Good| \geq (1-\epsilon)2^d$. Thus, for any $s \in Good$, $|\text{Ext}(\cdot, s)^{-1}(r)| = 2^{n-m}$. Thus, we have

$$\sum_{s \in Good} |\text{Ext}^{-1}(\cdot, s)(r)| \geq 2^{d+n-m-1}.$$

Further, for any $s' \in Bad$, $|\text{Ext}^{-1}(\cdot, s')(r)| \leq 2^n$, and hence

$$\sum_{s' \in Bad} |\text{Ext}^{-1}(\cdot, s')(r)| \leq \epsilon 2^{d+n} < 2^{d+n-1.5m}.$$

Thus $|\cup_{s' \in bad} \text{Ext}^{-1}(\cdot, s')(r)| < 2^{-0.5m}|\text{Ext}^{-1}(r)|$. It now follows that

$$|\mathcal{S}(r) - \mathcal{D}_r| \leq 2^{-0.4m}$$

□

We note that ϵ must be $o(2^{-m})$ for the above sampling procedure to work with low enough error. However, this would require a seed length of $d = O(m^2)$ (by Theorem 2.1.5). For each step of the alternating extraction protocol the seed length then goes down by a quadratic factor, which is insufficient for our application.

To get past this difficulty, we construct a new strong linear seeded extractor for high min-entropy sources with the seed length close to the output length with the property that the size of the pre-image of any output is the same for any fixing of the seed. Algorithm 11 provides this construction.

Parameters and Subroutines:

1. Let $\delta > 0$ be any constant. Let $d = n^\delta$. Let $d = d_1 + d_2$, where $d_1 = n^{\delta_1}$, $\delta > 10\delta_1$. Let $m = d/2$.

2. Let Samp : $\{0,1\}^{d_1} \to [n]^t$, $t = d_2$, be an (μ, θ, γ) averaging sampler with distinct samples, such that $\mu = \frac{(\delta - 2\tau)}{\log(1/\tau)}$, $\theta = \frac{\tau}{\log(1/\tau)}$ and $\gamma = 2^{-\Omega(d_1)}$, $\tau = 0.05$.

3. Let IP : $\{0,1\}^{d_2} \times \{0,1\}^{d_2} \to \{0,1\}^{\frac{d}{2}}$ be the strong 2-source extractor from Theorem 2.5.3.

Algorithm 11: iExt(x,s)
Input: Bit strings x, s of length n, d respectively.
Output: A bit string of length m.

1 Let $s_1 = \text{Slice}(s, d_1)$. Let s_2 be the remaining d_2 bits of s.
2 Let $T = \text{Samp}(s_1) \subset [n]$. Let $x_1 = x_{\{T\}}$.
3 Output $\text{IP}(x_1, s_2)$.

Informally the construction of iExt is as follows. Given a uniform seed $\mathbf{S}$, we use a slice $\mathbf{S}_1$ of $\mathbf{S}$ to sample co-ordinates from the weak source $\mathbf{X}$, and then apply a strong 2-source extractor (based on the inner product function) to the source $\mathbf{X}_1$ (which is the projection of $\mathbf{X}$ to the sampled co-ordinates) and the remaining bits $\mathbf{S}_2$ of $\mathbf{S}$ to extract $\frac{d}{2}$ uniform bits.

The correctness of this procedure relies on the fact that by pseudorandomly sampling co-ordinates of $\mathbf{X}$ and projecting $\mathbf{X}$ to these co-ordinates, the min-entropy rate is roughly the preserved for most choices of the uniform seed [Zuc97, Vad04, Li12a]. Thus, we can fix $\mathbf{S}_1$, and the strong two-source extractor IP now receives two independent inputs $\mathbf{S}_2$ and $\mathbf{X}_2$ with almost full min-entropy. Thus, the output is close to uniform. Further we show that the number of linear constraints on the source $\mathbf{X}$ is the same for any fixing of the seed. This allows us to show that size of the pre-image of any particular output is the same for any fixing of the seed. We now formally prove these ideas.

We need the following theorem proved by Vadhan [Vad04].

Theorem 12.4.4 ([Vad04])**.** *Let* $1 \geq \delta \geq 3\tau > 0$. *Let* $\text{Samp} : \{0,1\}^r \to [n]^t$ *be an* (μ, θ, γ) *averaging sampler with distinct samples, such that* $\mu = \frac{(\delta - 2\tau)}{\log(1/\tau)}$ *and* $\theta = \frac{\tau}{\log(1/\tau)}$. *If* $\mathbf{X}$ *is a* $(n, \delta n)$ *source, then the random variable* $(\mathbf{U}_r, \mathbf{X}_{\{Samp(\mathbf{U}_r)\}})$ *is* $(\gamma + 2^{-\Omega(\tau n)})$*-close to* $(\mathbf{U}_r, W)$ *where for every* $a \in \{0,1\}^r$ *, the random variable* $W|\mathbf{U}_r = a$ *is a* $(t, (\delta - 3\tau)t)$*-source.*

Lemma 12.4.5. *Let* iExt *be the function computed by Algorithm* 11. *If* $\mathbf{X}$ *is a* $(n, 0.9n)$ *source and* $\mathbf{S}$ *is an independent uniform seed on* $\{0,1\}^d$*, then the following holds:*

$$|\text{iExt}(\mathbf{X}, S), S - \mathbf{U}_m, S| < 2^{-n^{\Omega(1)}}.$$

Further for any $r \in \{0,1\}^m$ and any $s \in \{0,1\}^d$, $|\mathrm{iExt}(\cdot, s)^{-1}(r)| = 2^{n-m}$.

Proof. Using Theorem 12.4.4, it follows that $\mathbf{X}_1$ is $2^{-n^{\Omega(1)}}$-close to a source with min-entropy at least $0.8n$ for any fixing of $\mathbf{S}_1$. Further, we note that after fixing $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{X}_1$ are independent sources. We now think of $\mathbf{X}_1, \mathbf{S}_2$ as sources in $\{0,1\}^{d_2+1}$ by appending a 1 to both the sources, so that $\mathbf{S}_2 \neq \vec{0}$, and then apply the inner product map. This results in an entropy loss of only 1. It now follows by Theorem 2.5.3 that

$$|\mathrm{iExt}(\mathbf{X}, S), S - \mathbf{U}_m, S| < 2^{-n^{\Omega(1)}}.$$

It is easy to see that for any fixing of the seed $\mathbf{S} = s$, $\mathrm{iExt}(\cdot, s)$ is a linear map. Let $\mathbf{X}$ be uniform on n bits. We note that for any fixing of $\mathbf{S}_2 = s_2$, $\mathbf{X}_1$ lies in a subspace of dimension $d_2 - m$ over F_2. Further, the bits outside T have no restrictions placed on them. Thus the size of $\mathrm{iExt}(\cdot, s)^{-1}(r)$ is exactly $2^{d_2-m+n-d_2} = 2^{n-m}$. This completes the proof of the lemma. □

Based on the above lemma, we construct an efficient procedure for sampling uniformly from the pre-image of the function iExt.

Claim 12.4.6. *Let* $\mathrm{iExt} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be the function computed by Algorithm* 11*. Then there exists a polynomial time algorithm* Samp_1 *that takes as input $r \in \{0,1\}^m$, and samples from a distribution that is uniform on* $\mathrm{iExt}^{-1}(r)$*.*

Proof. It follows by Lemma 12.4.5 that for any fixing of the seed s, the size of the set $\mathrm{iExt}(\cdot, s)^{-1}(r)$ is exactly 2^{n-m}. Thus we can use the following strategy: (a) Sample $s \sim \mathbf{U}_d$ (b) Sample x uniformly random from the subspace $\mathrm{iExt}(\cdot, s)^{-1}(r)$ (c) Output (x, s). It follows that each element in $\mathrm{iExt}^{-1}(r)$ is picked with probability exactly $\frac{1}{2^d} \cdot \frac{1}{2^{n-m}}$. Thus the output of our sampling procedure is indeed uniform on $\mathrm{iExt}^{-1}(r)$. □

12.4.2 A Modified Construction of the Seedless $(2, t)$-Non-Malleable Extractor

We first describe the high level ideas involved in modifying the construction of nmExt (Algorithm 3), before presenting the formal construction.

- We use the linear seeded extractor iExt (Algorithm 11) for any seeded extractor used in the construction of nmExt.

- Next we divide the sources $\mathbf{X}$ and $\mathbf{Y}$ into blocks of size $n^{1-\delta}$ respectively for a small constant δ. Since each of $\mathbf{X}$ and $\mathbf{Y}$ have almost full min-entropy, we now have two block sources, where each block has almost full min-entropy conditioned on the previous blocks. The idea is to use new blocks of $\mathbf{X}$ and $\mathbf{Y}$ for each round of alternating extraction in nmExt.

  To implement this however, we need some care. Recall that the alternating extraction protocol is run for two rounds between either $\mathbf{X}$ and $\mathbf{Q}_h$, or $\mathbf{X}$ and $\overline{\mathbf{Q}}_h$ in the function 2laExt. The idea now is to run these two of alternating extraction by dividing $\mathbf{Q}_h$ into two blocks, and using two new partitions of $\mathbf{X}$ (each round being run by using a block from either $\mathbf{X}$ or $\mathbf{Q}_h$). Now to generate these $\mathbf{Q}_h$'s, we use a $O(t)$ blocks of $\mathbf{Y}$, and for each block apply the strong seeded extractor iExt, using as seed the output of the alternating extraction from the previous step, and finally concatenate the outputs. This works because these $O(t)$ blocks form a block source, and using the same seed to extract from all the blocks is a well known technique of extracting from block sources.

- By appropriate setting of the lengths of the seeds in the alternating extraction, we ensure that each block of $\mathbf{X}$ and $\mathbf{Y}$ still has min-entropy rate $1 - o(1)$ even after fixing all the intermediate seeds, the random variables $\mathbf{Q}_h, \overline{\mathbf{Q}}_h$ and their tampered versions. This can be ensured since each of these variables are of length at most n^{δ_1} for some small constant δ_1, and the number of adversaries is also $n^{\Omega(1)}$).

- The above modification is almost sufficient for us to successfully sample from the pre-image of any output. One final modification is to use a specific error correcting code (the Reed-Solomon

code over a field of size $n+1$ with characteristic 2) in the initial step of the construction, when we encode the sources and sample bits from it. We give some intuition as to why this is necessary. Since we are using linear seeded extractors in the alternating extraction, by fixing the seeds we impose linear restrictions on the blocks of $\mathbf{X}$ and $\mathbf{Y}$. Now, if we fix the output of the initial sampling step (the random variable $\mathbf{Z}$ in Algorithm 3), we are imposing more linear constraints on the blocks (assuming we are using a linear code). Now, it is not clear if the constraints imposed by the linear seeded extractor is independent from the constraints imposed by $\mathbf{Z}$, and thus for different fixings of the $\mathbf{Z}$ and the seeds the size of the pre-image of any output of the non-malleable extractor may be different.

To get past this difficulty, our idea is to first partition $\mathbf{X}$ and $\mathbf{Y}$ into slightly smaller blocks (which does not affect the correctness of the extractor) such that at least half of the blocks are unused by the alternating extraction steps. Now, we show that by using the Reed-Solomon code over $\mathbb{F} = \mathbb{F}_{2^{\log(n+1)}}$ to encode the sources, fixing $\mathbf{Z}$ imposes linear constraints involving the variables from these unused blocks, and we show that this is sufficient to argue that it is linearly independent of the restrictions imposed by the alternating extraction part. We provide complete details of the sampling algorithms in Section 12.4.3.

We now proceed to present the extractor construction. Recall that if $\mathbf{Z}_a, \mathbf{Z}_{a+1}, \ldots, \mathbf{Z}_b$ are random variables, we use $\mathbf{Z}_{[a,b]}$ to denote the random variable $\mathbf{Z}_a, \ldots, \mathbf{Z}_b$.

Subroutines and Parameters (used by Algorithm 12, Algorithm 13, Algorithm 14)

1. Let γ be a small enough constant and C a large one. Let $t = n^{\gamma/C}$.

2. Let $n_1 = n^{\beta_1}$, $\beta_1 = 10\gamma$. Let $n_2 = n - n_1$. Let $\mathrm{IP}_1 : \{0,1\}^{n_1} \times \{0,1\}^{n_1} \to \{0,1\}^{n_3}$, $n_3 = \frac{n_1}{10}$ be the strong two-source extractor from Theorem 2.5.3.

3. Let $\mathbb{F}$ be the finite field $\mathbb{F}_{2^{\log(n+1)}}$. Let $n_4 = \frac{n_2}{\log(n+1)}$. Let $\mathrm{RS} : \mathbb{F}^{n_4} \to \mathbb{F}^n$ be the Reed-Solomon code encoding n_4 symbols of $\mathbb{F}$ to n symbols in $\mathbb{F}$ (we overload the use of RS, using it to

denote both the code and the encoder). Thus RS is a $[n, n_4, n - n_4 + 1]_n$ error correcting code.

4. Let $\text{Samp} : \{0,1\}^{n_3} \to [n]^{n_5}$ be a $(\mu, \frac{1}{10}, 2^{-n^{\Omega(1)}})$ averaging sampler with distinct samples. By using the strong seeded extractor from Theorem 2.1.2, we can set $n_5 = n^{\beta_2}$, $\beta_2 < \beta_1/2$.

5. Let $\ell = 2(n_1 + n_5 \log n) < 4n^{\beta_1}$. Thus $\ell \leq n^{11\gamma}$.

6. Let $n_6 = 50Ct\ell$. Let $\text{IP}_2 : \{0,1\}^{n_6} \times \{0,1\}^{n_6} \to \{0,1\}^{2n_q}$, $n_q = 10Ct\ell$, be the strong two-source extractor from Theorem 2.5.3.

7. Let $n_7 = n - n_1 - n_6$. Let $n_x = \frac{n_7}{8\ell}$. Let $n_y = \frac{n_7}{16Ct\ell}$. Thus $n_x, n_y \geq n^{1-15\gamma}$.

8. Let $d_1 = 80\ell$.

9. Let $\text{iExt}_1 : \{0,1\}^{n_x} \times \{0,1\}^{d_1} \to \{0,1\}^{d_2}$, $d_2 = 40\ell$, be the extractor computed by Algorithm 11.

10. Let $\text{iExt}_2 : \{0,1\}^{n_q} \times \{0,1\}^{d_2} \to \{0,1\}^{d_3}$, $d_3 = 20\ell$, be the extractor computed by Algorithm 11.

11. Let $\text{iExt}_3 : \{0,1\}^{n_x} \times \{0,1\}^{d_3} \to \{0,1\}^{d_4}$, $d_4 = 10\ell$ be the extractor computed by Algorithm 11.

12. Let $\text{iExt}_4 : \{0,1\}^{n_y} \times \{0,1\}^{d_4} \to \{0,1\}^{d_5}$, $d_5 = 5\ell$, be the extractor computed by Algorithm 11.

13. Let $\text{Ext} : \{0,1\}^{4Ctn_y} \times \{0,1\}^{d_4} \to \{0,1\}^{2n_q}$ be defined in the following way. Let $v_1, \ldots, v_{4t}$ be strings, each of length n_y. Define $\text{Ext}(v_1 \circ \ldots \circ v_{4Ct}, s) = \text{iExt}_4(v_1, s) \circ \ldots \circ \text{iExt}_4(v_{4Ct}, s)$.

Theorem 12.4.7. *Let* inmExt *be the function computed by Algorithm* 13. *Then* inmExt *is a seedless* $(2, t)$*-non-malleable extractor with error* $2^{-n^{\Omega(1)}}$.

Algorithm 12: inmExt(x,y)

Input: Bit strings x, y, each of length n.
Output: A bit string of length m.

Let $x_1 = \text{Slice}(x, n_1)$, $y_1 = \text{Slice}(y, n_1)$. Compute $\nu = \text{IP}_1(x, y)$.
Let x_2, y_2 be n_2 length strings formed by cutting x_1, y_1 from x, y respectively.
Let $T = \text{Samp}(\nu) \subset [n]$.
Interpret x_2, y_2 as elements in $\mathbb{F}^{n_4}$.
Let $\overline{x}_2 = \text{RS}(x_2), \overline{y}_2 = \text{RS}(y_2)$.
Let $\overline{x}_1 = (\overline{x}_2)_{\{T\}}, \overline{y}_1 = (\overline{y}_2)_{\{T\}}$, interpreting $\overline{x}_2, \overline{y}_2 \in \mathbb{F}^n$.
Let $z = x_1 \circ \overline{x}_1 \circ y_1 \circ \overline{y}_1$, where z is interpreted as a binary string.
Interpret x_2, y_2 as binary strings.
Output $\text{inmExt}_1(x_2, y_2, z)$.

Algorithm 13: $\text{inmExt}_1(x_2, y_2, z)$

Let $x_3 = \text{Slice}(x_2, n_6), y_3 = \text{Slice}(y_2, n_6)$. Let w, v be the remaining parts of x_2, y_2 respectively.
Let $\text{IP}_2(x_3, y_3) = (q_{1,1}, q_{1,2})$, where each of $q_{1,1}, q_{1,2}$ is of length n_q.
Let $w_1, \ldots, w_{8\ell}$ be an equal sized partition of the string w into 8ℓ stings.
Let $v_1, \ldots, v_{16t\ell}$ be an equal sized partition of the string v into $16Ct\ell$ stings.
for $h = 1$ *to* ℓ **do**
  $(q_{h+1,1}, q_{h+1,2}) = \text{2ilaExt}(v_{[8C(h-1)t+1,8Cht]}, w_{[4h-3,4h]}, q_{h,1}, q_{h,2}, h, z_{\{h\}})$
end
Ouput $(q_{\ell+1,1}, q_{\ell+1,2})$.

The proof of the above theorem is essentially the same as that the construction in Section 11.3, and we omit it. The correctness of inmExt follows directly from the proof of Theorem 11.3.1, and the correctness of the extractor iExt (Lemma 12.4.5), the fact that by our choice of parameters each block of $\mathbf{X}$ and $\mathbf{Y}$ still has min-entropy rate at least 0.9 after appropriate conditioning of the intermediate random variables and their tampered versions, and the fact that using the RS in place of a binary error correcting code does not affect correctness of the procedure.

Algorithm 14: 2ilaExt$(v_{[8C(h-1)t+1,8Cht]}, w_{[4h-3,4h]}, q_{h,1}, q_{h,2}, h, b)$

Let $s_{h,1} = \text{Slice}(q_{h,1}, d_1)$, $r_{h,1} = \text{Ext}_1(w_{4h-3}, s_{h,1})$, $s_{h,2} = \text{Ext}_2(q_{h,2}, r_{h,1})$, $r_{h,2} = \text{Ext}_3(w_{4h-2}, s_{h,2})$.
if $b = 0$ **then**
  Let $r_h = \text{Slice}(r_{h,1}, d_4)$.
else
  Let $r_h = r_{h,2}$
end
Let $\text{Ext}(v_{[8C(h-1)t+1,8(h-1)t+4Ct]}, r_h) = (\overline{q}_{h,1}, \overline{q}_{h,2})$, where both $\overline{q}_{h,1}, \overline{q}_{h,2}$ are of length n_q.
Let $\overline{s}_{h,1} = \text{Slice}(\overline{q}_{h,1}, d_1)$, $\overline{r}_{h,1} = \text{Ext}_1(w_{4h-1}, \overline{s}_{h,1})$, $\overline{s}_{h,2} = \text{Ext}_2(\overline{q}_{h,2}, \overline{r}_{h,1})$, $\overline{r}_{h,2} = \text{Ext}_3(w_{4h}, \overline{s}_{h,2})$.
if $b = 0$ **then**
  Let $\overline{r}_h = \overline{r}_{h,2}$.
else
  Let $\overline{r}_h = \text{Slice}(\overline{r}_{h,1}, d_4)$.
end
Let $\text{Ext}(v_{[8C(h-1)t+4Ct+1,8Cht]}, r_h) = (q_{h+1,1}, q_{h+1,2})$, where both $q_{h+1,1}, q_{h+1,2}$ are of length n_q.
Ouput $(q_{h+1,1}, q_{h+1,2})$.

12.4.3 Efficiently Sampling from the Pre-Image of inmExt

Since the construction of the non-malleable extractor inmExt (Algorithm 12, Algorithm 13, Algorithm 14) is composed of various sub-parts and sub-functions, we first argue about the invertibility of these parts and then show a way to compose these sampling procedure to sample almost uniformly from the pre-image of inmExt. We refer to all the variables, sub-routines and notations introduced in these algorithms while developing the sampling procedures. Unless we state otherwise, by a subspace we mean a subspace over $\mathbb{F}_2$.

We first show how to sample uniformly from the pre-image of 2ilaExt (Algorithm 14), since it is a crucial sub-part of inmExt. We have the following claim.

Claim 12.4.8. *For any fixing of the variables* $\{s_{1,i}, r_{1,i}, \overline{s}_{1,i}, \overline{r}_{1,i} : i \in \{1,2\}\}$*, and any* $b \in \{0,1\}$

define the set:

$$\text{2ilaExt}^{-1}(q_{2,1}, q_{2,2}) = \{(x_3, y_3, v_{[1,4Ct]}, w_{[1,4]}) \in \{0,1\}^{2n_6+4Ctn_y+4n_x} : \\ \text{2ilaExt}(v_{[1,4Ct]}, w_{[1,4]}, q_{1,1}, q_{1,2}, b) = (q_{2,1}, q_{2,2})\}$$

There exists an efficient algorithm Samp_2 *that takes as input* $q_{2,1}, q_{2,2}, b, \{s_{1,i}, r_{1,i}, \overline{s}_{1,i}, \overline{r}_{1,i} : i \in \{1,2\}\}$*, and samples uniformly from* $\text{2ilaExt}^{-1}(q_{2,1}, q_{2,2})$*.*

Further, the set $\text{2ilaExt}^{-1}(q_{2,1}, q_{2,2})$ *is a subspace over* $\mathbb{F}_2$ *of dimension* d_1*, and its size does not depend on the inputs to* Samp_2*.*

Proof. The general idea is that by fixing the seeds in the alternating extraction, each block of w takes values independent of the fixing of the other blocks of w and the $q_{i,j}$'s, and similarly the $q_{i,j}$'s takes values independent of each other and the blocks of w. We now formally prove this intuition.

Since, $s_{1,1}$ is a slice of $q_{1,1}$ it follows that $q_{1,1}$ is restricted to the subspace of size $2^{n_q-d_1}$. Since $r_{1,1} = \text{iExt}_1(w_1, s_{1,1})$, it follows that w_1 is restricted to the set $\text{iExt}_1(\cdot, s_{1,1})^{-1}(r_{1,1})$. Further, it follows by Lemma 12.4.5 that this is a subspace of size $2^{n_x-d_2}$. Similar arguments show that $q_{1,2}$ is restricted to the subspace of dimension $2^{n_q-d_3}$, and w_2 is restricted to a subspace of dimension $2^{n_x-d_4}$. Further, we note that each of these variables have no correlation.

By repeating this argument for the next two rounds of alternating extraction, it follows that $\overline{q}_{1,1}$ is restricted to a subspace of size $2^{n_q-d_1}$, w_3 is restricted to a subspace of size $2^{n_x-d_2}$, $\overline{q}_{1,2}$ is restricted to a subspace of size $2^{n_q-d_3}$, and w_4 is restricted to a subspace of size $2^{n_x-d_4}$.

Further since $(q_{2,1}, q_{2,2}) = \text{Ext}(v_{[4Ct+1,8t]}, r_1) = \text{iExt}_4(v_{4Ct+1}, r_1) \circ \ldots \circ \text{iExt}_4(v_{8Ct}, r_1)$, it follows by an application of Lemma 12.4.5 that for any fixed $q_{2,1}$, $v_{[4Ct+1,6t]}$ is restricted to a subspace of size $2^{2Ct(n_y-d_5)}$. A similar argument shows that for any fixed $q_{2,2}$, $v_{[6Ct+1,8Ct]}$ is restricted to a subspace of size $2^{2Ct(n_y-d_5)}$.

Finally, since $\text{IP}_1(x_3, y_3) = (q_{1,1}, q_{1,2})$, it follows that for any fixed $x_3, q_{1,1}, q_{1,2}$, the variable y_3 lies in a subspace of size $2^{n_6-\log(2n_q)}$ since by fixing the variables $x_3, q_{1,1}, q_{1,2}$, we are restricting

y_3 to a subspace of dimension $\left(\frac{n_6}{\log(2n_q)} - 1\right)$ over the field $\mathbb{F}_{2^{\log(2n_q)}}$.

It is clear from the arguments that we did not use any specific values of the inputs given to the algorithm Samp_1 (including the value of the bit b) to argue about the size of $\text{2ilaExt}^{-1}(q_{2,1}, q_{2,2})$. Also note that each of $x_3, y_3, v_{[1,4Ct]}, w_{[1,4]}$ is restricted to some subspace. Since $\text{2ilaExt}^{-1}(q_{2,1}, q_{2,2})$ is the cartesian product of these subspaces, it follows that it is a subspace over $\mathbb{F}_2$. Thus the lemma now follows since we can efficiently sample from a given subspace. □

Using arguments very similar to the above claim, we obtain the following result.

Claim 12.4.9. *For any $h \in \{2, \ldots, \ell\}$, any fixing of the variables $\{s_{h,i}, r_{h,i}, \overline{s}_{h,i}, \overline{r}_{h,i} : i \in \{1,2\}\}$, and any $b \in \{0,1\}$ define the set:*

$$\text{2ilaExt}^{-1}(q_{h+1,1}, q_{h+1,2}) = \{(v_{[8C(h-1)t-4Ct+1,8C(h-1)t+4Ct]}, w_{[4h-3,4h]}) \in \{0,1\}^{8Ctn_y+4n_x} : \\ \text{2ilaExt}(v_{[8C(h-1)t+1,8Cht]}, w_{[4h-3,4h]}, q_{1,1}, q_{1,2}, b) = (q_{h+1,1}, q_{h+1,2})\}.$$

There exists an efficient algorithm Samp_{h+1} that takes as input $q_{h+1,1}, q_{h+1,2}, b, \{s_{h,i}, r_{h,i}, \overline{s}_{h,i}, \overline{r}_{h,i} : i \in \{1,2\}\}$, and samples uniformly from $\text{2ilaExt}^{-1}(q_{h+1,1}, q_{h+1,2})$.

Further, $\text{2ilaExt}^{-1}(q_{h+1,1}, q_{h+1,2})$ is a subspace over $\mathbb{F}_2$, and its size does not depend on the inputs to Samp_{h+1}.

□

We now show a way of efficiently sampling from the pre-image of the function inmExt_1 (Algorithm 13).

Claim 12.4.10. *For any string $\alpha \in \{0,1\}^\ell$, and any fixing of the variables $\{s_{h,i}, r_{h,i}, \overline{s}_{h,i}, \overline{r}_{h,i} : h \in [\ell], i \in \{1,2\}\}$ define the set:*

$$\text{inmExt}_1^{-1}(q_{\ell+1,1}, q_{\ell+1,2}) = \{(x_2, y_2) \in \{0,1\}^{2n_2} : \text{inmExt}_1(x_2, y_2, \alpha) = (q_{\ell+1,1}, q_{\ell+1,2})\}.$$

There exists an efficient algorithm Samp_{nm_1} that takes as input $\{s_{h,i}, r_{h,i}, \overline{s}_{h,i}, \overline{r}_{h,i} : h \in [\ell], i \in$

$\{1,2\}\}, \alpha, q_{\ell+1,1}, q_{\ell+1,2}$, *and samples uniformly from* $\mathrm{inmExt}_1^{-1}(q_{\ell+1,1}, q_{\ell+1,2})$.

Further, $\mathrm{inmExt}_1^{-1}(q_{\ell+1,1}, q_{\ell+1,2})$ *is a subspace over* $\mathbb{F}_2$*, and its size does not depend on the inputs to* Samp_{nm_1}.

Proof. We observe that once we fix all the seeds $\{s_{h,i}, r_{h,i}, \overline{s}_{h,i}, \overline{r}_{h,i} : h \in [\ell], i \in \{1,2\}\}$, for different $h \in [\ell]$, the blocks $(v_{[8C(h-1)t-4Ct+1,8C(h-1)t+4Ct]}, w_{[4h-3,4h]})$ can be sampled independently. Thus, by using the algorithms $\{\mathrm{Samp}_{h+1} : h \in \ell\}$ from Claim 12.4.8 and Claim 12.4.9, we sample the variable $x_3, y_3, w_{[1,4]}, v_{[1,4Ct]}, \{v_{[8C(h-1)t-4Ct+1,8C(h-1)t+4Ct]}, w_{[4h-3,4h]} : h \in [\ell]\}$.

Finally, since $\mathrm{Ext}(v_{[8C(\ell-1)t+4Ct+1,8C\ell t]}, \overline{r}_\ell) = (q_{\ell+1,1}, q_{\ell+1,2})$, it follows by the arguments in Lemma 12.4.8, that the block $v_{[8C(\ell-1)t+4Ct+1,8C\ell t)]}$ is restricted to a subspace of size $2^{4Ct(n_y-d_5)}$. Thus, we can efficiently sample this block as well.

Further the variable $w_{[4\ell+1,8\ell]}$ is unused by the algorithm inmExt_1, and hence takes all values in $\{0,1\}^{4\ell n_x}$. Similarly the variable $v_{[8C\ell t+1,16C\ell t]}$ is unused by the algorithm inmExt_1 and hence takes all values in $\{0,1\}^{8Ct\ell}$. Thus, we sample these variables as uniform strings of the appropriate length.

Since x_2, y_2 are concatenations of the various blocks sampled above, we can indeed sample efficiently from a distribution uniform on $\{(x_2, y_2) \in \{0,1\}^{2n_2} : \mathrm{inmExt}(x, y, \alpha) = (q_{\ell+1,1}, q_{\ell+1,2})\}$. Further since by Claim 12.4.8 and Claim 12.4.9, the size of the pre-images of each of the blocks generated do not depend on the inputs (and is also a subspace), it follows that $2\mathrm{inmExt}_1^{-1}(q_{\ell+1,1}, q_{\ell+1,2})$ is a subspace, and its size does not depend on the inputs to Samp_{nm_1}. □

We now proceed to construct an algorithm to uniformly sample from the pre-image of any output of the function inmExt (Algorithm 12), which will yield the required efficient encoder for the resulting one-many non-malleable codes.

Claim 12.4.11. *For any fixing of the variable* $z = x_1 \circ \overline{x}_1 \circ y_1 \circ \overline{y}_1$ *and the variables* $\{s_{h,i}, r_{h,i}, \overline{s}_{h,i}, \overline{r}_{h,i} :$

$h \in [\ell], i \in \{1,2\}\}$, *define the set:*

$$\mathrm{inmExt}^{-1}(q_{\ell+1,1}, q_{\ell+1,2}) = \{(x,y) \in \{0,1\}^{2n} : \mathrm{inmExt}(x,y) = (q_{\ell+1,1}, q_{\ell+1,2})\}.$$

There exists an efficient algorithm Samp_{nm} *that takes as input* $\{s_{h,i}, r_{h,i}, \overline{s}_{h,i}, \overline{r}_{h,i} : h \in [\ell], i \in \{1,2\}\}, z, q_{\ell+1,1}, q_{\ell+1,2}$*, and samples uniformly from* $\mathrm{inmExt}^{-1}(q_{\ell+1,1}, q_{\ell+1,2})$.

Further, $\mathrm{inmExt}^{-1}(q_{\ell+1,1}, q_{\ell+1,2})$ *is a subspace over* $\mathbb{F}_2$*, and its size does not depend on the inputs to* Samp_{nm}.

Proof. We fix the variables x_1 and y_1. Let $T = \mathrm{Samp}(\nu) = \{t_1, \ldots, t_{n_5}\}$. We now think of x_2 as an element in $\mathbb{F}^{n_4}$, $\mathbb{F} = \mathbb{F}_{2^{\log(n+1)}}$. Let $x_2 = (x_{2,1}, \ldots, x_{2,n_4})$, where each $x_{2,i}$ is in $\mathbb{F}$. Recall that the $n_4 \times n$ generator matrix G of the code RS is the following:

$$G = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{n_4-1} & \alpha_2^{n_4-1} & \cdots & \alpha_n^{n_4-1} \end{pmatrix}$$

where $\alpha_1, \ldots, \alpha_n$ are distinct non-zero field elements of $\mathbb{F}$.

Let

$$G_T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \alpha_{t_1} & \alpha_{t_2} & \cdots & \alpha_{t_{n_5}} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{t_1}^{n_4-1} & \alpha_{t_2}^{n_4-1} & \cdots & \alpha_{t_{n_5}}^{n_4-1} \end{pmatrix}$$

Since $\overline{x}_1 = \mathrm{RS}(x_2)_{\{T\}}$, we have the following identity:

$$\begin{pmatrix} x_{2,1} & \cdots & x_{2,n_4} \end{pmatrix} G_T = \overline{x}_1 \tag{12.3}$$

Thus, for any fixing of $\overline{x}_1$, the variable x_2 is restricted to a subspace of dimension $(n_4 - n_5)$ over

the field $\mathbb{F}$.

Now, let $j \in [n_4]$ be such that $(x_{2,1}, \ldots, x_{2,j})$ is the string $(x_3, w_{[1,4\ell]})$, and $(x_{2,j+1}, \ldots, x_{2,n_4})$ is the string $w_{[4\ell+1,8\ell]}$. Clearly, $(n_4 - j)\log n = 4\ell n_x$, and thus by our choice of parameters it follows that $j = n_4 - \frac{4\ell n_x}{\log n} = \frac{n_4}{2} + \frac{n_6}{log(n+1)} < \frac{2n_4}{3} < n_4 - n_5$.

We further note since any $n_5 \times n_5$ sub-matrix of G_T has full rank (since it is the Vandermonde's matrix), it follows by the rank-nullity thorem that any $j \times n_5$ sub-matrix of G_T has null space of dimension exactly $j - n_5$. Thus for any $\lambda \in \mathbb{F}^{n_5}$, the equation:

$$\begin{pmatrix} x_{2,j+1} & \cdots & x_{2,n_4} \end{pmatrix} \begin{pmatrix} \alpha_{t_1}^{j} & \alpha_{t_2}^{j} & \cdots & \alpha_{t_{n_5}}^{j} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{t_1}^{n_4-1} & \alpha_{t_2}^{n_4-1} & \cdots & \alpha_{t_{n_5}}^{n_4-1} \end{pmatrix} = \overline{x}_1 + \lambda \tag{12.4}$$

has exactly $|\mathbb{F}|^{(j-n_5)}$ solution.

Thus, for any fixing of the variables, $x_{2,1}, \ldots, x_{2,j}$, equation (1) has exactly $|\mathbb{F}|^{j-n_5}$ solutions. In other words, for any fixing of $x_3, w_{[1,4\ell]}, \overline{x}_1$, the variable $w_{[4\ell+1,8\ell]}$ is restricted to a subspace, and the size of the subspace does not depend on the fixing of $x_3, w_{[1,4\ell]}, \overline{x}_1$. Using, a similar argument, we can show that for any fixing of $y_3, v_{[1,8Ct\ell]}, \overline{y}_1$, the variable $v_{[8Ct\ell+1,16Ct\ell]}$ is restricted to a subspace, and the size of the subspace does not depend on the fixing of $y_3, v_{[1,8Ct\ell]}, \overline{y}_1$.

Now consider any fixing of the variables $\{s_{h,i}, r_{h,i}, \overline{s}_{h,i}, \overline{r}_{h,i} : h \in [\ell], i \in \{1,2\}\}, z$. As proved in the Claim 12.4.10, we can efficiently sample the variables $x_3, w_{[1,4\ell]}, y_3, v_{[1,8Ct\ell]}$. By the above argument, the variables $v_{[4\ell+1,8\ell]}$ and $w_{[8Ct\ell+1,16Ct\ell]}$ now lie in a subspace, and hence we can efficiently sample these variables as well. Thus we have an efficient procedure Samp_{nm} for uniformly sampling (x, y) from the set $\mathrm{inmExt}^{-1}(q_{\ell+1,1}, q_{\ell+1,2})$.

It also follows by Claim 12.4.10, that the total size of the pre-image of the variables $x_3, w_{[1,4\ell]}, y_3, v_{[1,8Ct\ell]}$ does not depend on z or the variables $\{s_{h,i}, r_{h,i}, \overline{s}_{h,i}, \overline{r}_{h,i} : h \in [\ell], i \in \{1,2\}\}$. Further, for any fixing of $x_3, w_{[1,4\ell]}, y_3, v_{[1,8Ct\ell]}, z$, as argued above, the variables $v_{[4\ell+1,8\ell]}$ and $w_{[8Ct\ell+1,16Ct\ell]}$ now lie in a subspace, whose size does not depend on the fixed variables. Thus,

overall the size of the total pre-image of x, y does not depend on the inputs to Samp_{nm}. □

We now state the main result of this section.

Theorem 12.4.12. *There exists an efficient procedure that given an input* $(q_{\ell+1,1}, q_{\ell+1,2}) \in \{0,1\}^{n_q} \times \{0,1\}^{n_q}$*, samples uniformly from the set* $\{(x,y) : \text{inmExt}(x,y) = (q_{\ell+1,1}, q_{\ell+1,2})\}$.

Proof. We use the following simple strategy.

1. Uniformly sample the variables $z, \{s_{h,i}, r_{h,i}, \overline{s}_{h,i}, \overline{r}_{h,i} : h \in [\ell], i \in \{1,2\}\}$,
2. Use the variables sampled in Step (1) as input to the algorithm Samp_{nm} to sample (x,y).

The correctness of this procedure follows directly from Claim 12.4.11, since it was proved that for any fixing of the variables of Step 1, the size of pre-image of inmExt is the same. □

Bibliography

[ABN+92] Noga Alon, Jehoshua Bruck, Joseph Naor, Moni Naor, and Ron M. Roth. Construction of asymptotically good low-rate error-correcting codes through pseudo-random graphs. *IEEE Transactions on Information Theory*, 38:509–516, 1992.

[ACRT97] A. E. Andreev, A. E. F. Clementi, J. D. P. Rolim, and L. Trevisan. Weak random sources, hitting sets, and bpp simulations. In *Foundations of Computer Science, 1997. Proceedings., 38th Annual Symposium on*, pages 264–272, Oct 1997.

[ADJ+14] Divesh Aggarwal, Yevgeniy Dodis, Zahra Jafargholi, Eric Miles, and Leonid Reyzin. Amplifying privacy in privacy amplification. In *CRYPTO*, 2014.

[ADKO15] D. Aggarwal, Y. Dodis, T. Kazana, and M. Obremski. Non-malleable reductions and applications. To appear in STOC, 2015.

[ADL14] Divesh Aggarwal, Yevgeniy Dodis, and Shachar Lovett. Non-malleable codes from additive combinatorics. In *STOC*, 2014.

[AGM03] Noga Alon, Oded Goldreich, and Yishay Mansour. Almost k-wise independence versus k-wise independence. *Inf. Process. Lett.*, 88(3):107–110, 2003.

[AGM+14] Shashank Agrawal, Divya Gupta, Hemanta K. Maji, Omkant Pandey, and Manoj Prabhakaran. Explicit non-malleable codes resistant to permutations. Cryptology ePrint Archive, Report 2014/316, 2014.

[AHL15] Divesh Aggarwal, Kaave Hosseini, and Shachar Lovett. Affine-malleable extractors, spectrum doubling, and application to privacy amplification. Cryptology ePrint Archive, Report 2015/1094, 2015.

[AKO15] Divesh Aggarwal, Tomasz Kazana, and Maciej Obremski. Inception makes non-malleable codes stronger. Cryptology ePrint Archive, Report 2015/1013, 2015.

[AL93] Miklós Ajtai and Nathan Linial. The influence of large coalitions. *Combinatorica*, 13(2):129–145, 1993.

[Alo98] Noga Alon. The Shannon capacity of a union. *Combinatorica*, 18(3):301–310, 1998.

[AM86] Noga Alon and Wolfgang Maass. Meanders, Ramsey Theory and Lower Bounds for Branching Programs. In *IEEE Symposium on Foundations of Computer Science*, pages 410–417, 1986.

[AN93] Noga Alon and Moni Naor. Coin-flipping games immune against linear-sized coalitions. *SIAM J. Comput.*, 22(2):403–417, 1993.

[AS92] Noga Alon and Joel Spencer. *The Probabilistic Method*. John Wiley, 1992.

[Bar06] Boaz Barak. A Simple Explicit Construction of an $n^{\tilde{O}(\log n)}$-Ramsey Graph. Technical report, Citeseer, 2006.

[BBCM95] C. H. Bennett, G. Brassard, C. Crepeau, and U. M. Maurer. Generalized privacy amplification. *IEEE Transactions on Information Theory*, 41(6):1915–1923, Nov 1995.

[BBR88] C.H. Bennett, G. Brassard, and J.-M. Robert. Privacy amplification by public discussion. *SIAM Journal on Computing*, 17:210–229, 1988.

[BCG15] Joseph Bonneau, Jeremy Clark, and Steven Goldfeder. On bitcoin as a public randomness source. *IACR Cryptology ePrint Archive*, 2015:1015, 2015.

[BGJT14] Razvan Barbulescu, Pierrick Gaudry, Antoine Joux, and Emmanuel Thomé. A heuristic quasi-polynomial algorithm for discrete logarithm in finite fields of small characteristic. In *Advances in Cryptology - EUROCRYPT 2014 - 33rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Copenhagen, Denmark, May 11-15, 2014. Proceedings*, pages 1–16, 2014.

[BGK06] J. Bourgain, A. A. Glibichuk, and S. V. Konyagin. Estimates for the number of sums and products and for exponential sums in fields of prime order. *Journal of the London Mathematical Society*, 73:380–398, 4 2006.

[BIW06] Boaz Barak, Russell Impagliazzo, and Avi Wigderson. Extracting randomness using few independent sources. *SIAM J. Comput.*, 36(4):1095–1118, December 2006.

[BKS+10] Boaz Barak, Guy Kindler, Ronen Shaltiel, Benny Sudakov, and Avi Wigderson. Simulating independence: New constructions of condensers, Ramsey graphs, dispersers, and extractors. *J. ACM*, 57(4), 2010.

[BKT04] Jean Bourgain, Nets Katz, and Terence Tao. A sum-product estimate in finite fields, and applications. *Geometric and Functional Analysis GAFA*, 14(1):27–57, 2004.

[BL85] Michael Ben-Or and Nathan Linial. Collective coin flipping, robust voting schemes and minima of Banzhaf values. In *26th Annual Symposium on Foundations of Computer Science, Portland, Oregon, USA, 21-23 October 1985*, pages 408–416, 1985.

[Blu86] Manuel Blum. Independent unbiased coin flips from a correlated biased source finite state Markov chain. *Combinatorica*, 6(2):97–108, 1986.

[BN96] Ravi B. Boppana and Babu O. Narayanan. The biased coin problem. *SIAM J. Discrete Math.*, 9(1):29–36, 1996.

[Bou05a] J. Bourgain. Mordell's exponential sum estimate revisited. *Journal of the American Mathematical Society*, 18, No. 2 Apr.:477–499, 2005.

[Bou05b] J. Bourgain. More on the sum-product phenomenon in prime fields and its applications. *International Journal of Number Theory*, 01(01):1–32, 2005.

[Bou07] Jean Bourgain. On the construction of affine extractors. *GAFA Geometric And Functional Analysis*, 17(1):33–57, 2007.

[Bra10] Mark Braverman. Polylogarithmic independence fools AC^0 circuits. *J. ACM*, 57(5), 2010.

[BRSW12] Boaz Barak, Anup Rao, Ronen Shaltiel, and Avi Wigderson. 2-source dispersers for $n^{o(1)}$ entropy, and Ramsey graphs beating the Frankl-Wilson construction. *Annals of Mathematics*, 176(3):1483–1543, 2012. Preliminary version in STOC '06.

[BS89] Ravi Boppona and Joel Spencer. A useful elementary correlation inequality. *Journal of Combinatorial Theory, Series A*, 50(2):305 – 307, 1989.

[BS94] Antal Balog and Endre Szemerdi. A statistical theorem of set addition. *Combinatorica*, 14(3):263–268, 1994.

[BSZ11] Eli Ben-Sasson and Noga Zewi. From affine to two-source extractors via approximate duality. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, 2011.

[CCFP11] Hervé Chabanne, Gérard D. Cohen, Jean-Pierre Flori, and Alain Patey. Non-malleable codes from the wire-tap channel. *CoRR*, abs/1105.3879, 2011.

[CCP12] Hervé Chabanne, Gérard D. Cohen, and Alain Patey. Secure network coding and non-malleable codes: Protection against linear tampering. In *ISIT*, pages 2546–2550, 2012.

[CDF+08] Ronald Cramer, Yevgeniy Dodis, Serge Fehr, Carles Padró, and Daniel Wichs. Detection of algebraic manipulation with applications to robust secret sharing and fuzzy extractors. In *EUROCRYPT*, pages 471–488, 2008.

[CG88] Benny Chor and Oded Goldreich. Unbiased bits from sources of weak randomness and probabilistic communication complexity. *SIAM Journal on Computing*, 17(2):230–261, 1988.

[CG14a] Mahdi Cheraghchi and Venkatesan Guruswami. Capacity of non-malleable codes. In *ITCS*, pages 155–168, 2014.

[CG14b] Mahdi Cheraghchi and Venkatesan Guruswami. Non-malleable coding against bit-wise and split-state tampering. In *TCC*, pages 440–464, 2014.

[CGH+85] Benny Chor, Oded Goldreich, Johan Hastad, Joel Friedman, Steven Rudich, and Roman Smolensky. The bit extraction problem of t-resilient functions (preliminary version). In *26th Annual Symposium on Foundations of Computer Science, Portland, Oregon, USA, 21-23 October 1985*, pages 396–407, 1985.

[CGL16] Eshan Chattopadhyay, Vipul Goyal, and Xin Li. Non-malleable extractors and codes, with their many tampered extensions. In *STOC*, 2016.

[CKM11] Seung Geol Choi, Aggelos Kiayias, and Tal Malkin. *Advances in Cryptology – ASIACRYPT 2011: 17th International Conference on the Theory and Application of Cryptology and Information Security, Seoul, South Korea, December 4-8, 2011. Proceedings*, chapter BiTR: Built-in Tamper Resilience, pages 740–758. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[CKOR10] N. Chandran, B. Kanukurthi, R. Ostrovsky, and L. Reyzin. Privacy amplification with asymptotically optimal entropy loss. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing*, pages 785–794, 2010.

[CL16a] Eshan Chattopadhyay and Xin Li. Explicit non-malleable extractors, multi-source extractors and almost optimal privacy amplification protocols. *Electronic Colloquium on Computational Complexity (ECCC)*, 2016.

[CL16b] Eshan Chattopadhyay and Xin Li. Extractors for sumset sources. In *STOC*, 2016.

[CM06] Antonio Cafure and Guillermo Matera. Improved explicit estimates on the number of solutions of equations over a finite field. *Finite Fields Appl.*, 12(2):155–185, April 2006.

[Coh15a] Gil Cohen. Local correlation breakers and applications to three-source extractors and mergers. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, 2015.

[Coh15b] Gil Cohen. Local correlation breakers and applications to three-source extractors and mergers. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, 2015.

[Coh16a] Gil Cohen. Non-malleable extractors - new tools and improved constructions. In *CCC*, 2016.

[Coh16b] Gil Cohen. Non-malleable extractors with logarithmic seeds. Technical Report TR16-030, ECCC, 2016.

[Coh16c] Gil Cohen. Two-source dispersers for polylogarithmic entropy and improved Ramsey graphs. In *STOC*, 2016.

[CRS12] Gil Cohen, Ran Raz, and Gil Segev. Non-malleable extractors with short seeds and applications to privacy amplification. In *IEEE Conference on Computational Complexity*, pages 298–308, 2012.

[CRS14] Gil Cohen, Ran Raz, and Gil Segev. Non-malleable extractors with short seeds and applications to privacy amplification. *SIAM Journal on Computing*, 43(2):450–476, 2014.

[CS09] Mahdi Cheraghchi and Amin Shokrollahi. Almost-uniform sampling of points on high-dimensional algebraic varieties. In *STACS*, pages 277–288, 2009.

[CS16] Gil Cohen and Leonard Schulman. Extractors for near logarithmic min-entropy. Technical Report TR16-014, ECCC, 2016.

[CZ14] Eshan Chattopadhyay and David Zuckerman. Non-malleable codes against constant split-state tampering. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, pages 306–315, 2014.

[CZ16a] Eshan Chattopadhyay and David Zuckerman. Explicit two-source extractors and resilient functions. In *STOC*, 2016.

[CZ16b] Eshan Chattopadhyay and David Zuckerman. New extractors for interleaved sources. In *CCC*, 2016.

[DGJ+10] Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A. Servedio, and Emanuele Viola. Bounded independence fools halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010.

[DK11] Evgeny Demenkov and Alexander S. Kulikov. An elementary proof of a 3n - o(n) lower bound on the circuit complexity of affine dispersers. In *Proceedings of the 36th International Conference on Mathematical Foundations of Computer Science*, MFCS'11, pages 256–265, Berlin, Heidelberg, 2011. Springer-Verlag.

[DKO13] Stefan Dziembowski, Tomasz Kazana, and Maciej Obremski. Non-malleable codes from two-source extractors. In *CRYPTO (2)*, pages 239–257, 2013.

[DKRS06] Y. Dodis, J. Katz, L. Reyzin, and A. Smith. Robust fuzzy extractors and authenticated key agreement from close secrets. In *Advances in Cryptology — CRYPTO '06, 26th Annual International Cryptology Conference, Proceedings*, pages 232–250, 2006.

[DKSS09] Zeev Dvir, Swastik Kopparty, Shubhangi Saraf, and Madhu Sudan. Extensions to the method of multiplicities, with applications to Kakeya sets and mergers. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 181–190, 2009.

[DL12] Zeev Dvir and Shachar Lovett. Subspace evasive sets. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 351–358. ACM, 2012.

[DLWZ14] Yevgeniy Dodis, Xin Li, Trevor D. Wooley, and David Zuckerman. Privacy amplification and non-malleable extractors via character sums. *SIAM Journal on Computing*, 43(2):800–830, 2014.

[DO03] Y. Dodis and R. Oliveira. On extracting private randomness over a public channel. In *RANDOM*, pages 252–263, 2003.

[Dod06] Yevgeniy Dodis. Fault-tolerant leader election and collective coin-flipping in the full information model, 2006.

[DOPS04] Y. Dodis, Shien Jin Ong, M. Prabhakaran, and A. Sahai. On the (im)possibility of cryptography with imperfect randomness. In *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*, pages 196–205, Oct 2004.

[DORS08] Y. Dodis, R. Ostrovsky, L. Reyzin, and A. Smith. Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. *SIAM Journal on Computing*, 38:97–139, 2008.

[DP07] Stefan Dziembowski and Krzysztof Pietrzak. Intrusion-resilient secret sharing. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 227–237, Washington, DC, USA, 2007. IEEE Computer Society.

[DPW10] Stefan Dziembowski, Krzysztof Pietrzak, and Daniel Wichs. Non-malleable codes. In *ICS*, pages 434–452, 2010.

[DW09] Yevgeniy Dodis and Daniel Wichs. Non-malleable extractors and symmetric key cryptography from weak secrets. In *STOC*, pages 601–610, 2009.

[DY13] Yevgeniy Dodis and Yu Yu. Overcoming weak expectations. In *10th Theory of Cryptography Conference*, 2013.

[Fei99] Uriel Feige. Noncryptographic selection protocols. In *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science*, pages 142–153, 1999.

[FGHK15] Magnus Gausdal Find, Alexander Golovnev, Edward Hirsch, and Alexander Kulikov. A better-than-3n lower bound for the circuit complexity of an explicit function. Technical Report TR15-166, ECCC, 2015.

[FMNV14] Sebastian Faust, Pratyay Mukherjee, Jesper Buus Nielsen, and Daniele Venturi. Continuous non-malleable codes. In *TCC*, pages 465–488, 2014.

[FMVW13] Sebastian Faust, Pratyay Mukherjee, Daniele Venturi, and Daniel Wichs. Efficient non-malleable codes and key-derivation for poly-size tampering circuits. *IACR Cryptology ePrint Archive*, 2013:702, 2013.

[FW81] P. Frankl and R.M. Wilson. Intersection theorems with geometric consequences. *Combinatorica*, 1(4):357–368, 1981.

[GI02] Venkatesan Guruswami and Piotr Indyk. Near-optimal linear-time codes for unique decoding and new list-decodable codes over smaller alphabets. In *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing*, STOC '02, pages 812–821, New York, NY, USA, 2002. ACM.

[Gop14] Parikshit Gopalan. Constructing Ramsey graphs from boolean function representations. *Combinatorica*, 34(2):173–206, 2014.

[Gow98] W. T. Gowers. A new proof of szemeredi's theorem for arithmetic progressions of length four. *Geometric and Functional Analysis GAFA*, 8(3):529–551, 1998.

[GR08] Ariel Gabizon and Ran Raz. Deterministic extractors for affine sources over large fields. *Combinatorica*, 28(4):415–440, 2008.

[Gro00] Vince Grolmusz. Low rank co-diagonal matrices and Ramsey graphs. *Electr. J. Comb.*, 7, 2000.

[GRS06] Ariel Gabizon, Ran Raz, and Ronen Shaltiel. Deterministic extractors for bit-fixing sources by obtaining an independent seed. *SIAM J. Comput.*, 36(4):1072–1094, 2006.

[GSV05] S. Goldwasser, M. Sudan, and V. Vaikuntanathan. Distributed computing with imperfect randomness. In P. Fraigniaud, editor, *Proceedings of the 19th International Symposium on Distributed Computing DISC 2005*, volume 3724 of *Lecture Notes in Computer Science*, pages 288–302. Springer, 2005.

[Gur03] Venkatesan Guruswami. List decoding from erasures: bounds and code constructions. *IEEE Transactions on Information Theory*, 49(11):2826–2833, 2003.

[Gur04a] Venkatesan Guruswami. Better extractors for better codes? In *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, STOC '04, pages 436–444, New York, NY, USA, 2004. ACM.

[Gur04b] Venkatesan Guruswami. *List Decoding of Error-Correcting Codes (Winning Thesis of the 2002 ACM Doctoral Dissertation Competition)*, volume 3282 of *Lecture Notes in Computer Science*. Springer, 2004.

[Gur11] Venkatesan Guruswami. Linear-algebraic list decoding of folded Reed-Solomon codes. In *Computational Complexity (CCC), 2011 IEEE 26th Annual Conference on*, pages 77–85. IEEE, 2011.

[GUV09] Venkatesan Guruswami, Christopher Umans, and Salil P. Vadhan. Unbalanced expanders and randomness extractors from Parvaresh–Vardy codes. *J. ACM*, 56(4), 2009.

[HW98] Ming-Deh A. Huang and Yiu-Chung Wong. An algorithm for approximate counting of points on algebraic sets over finite fields. In *ANTS*, pages 514–527, 1998.

[Jan90] Svante Janson. Poisson approximation for large deviations. *Random Structures & Algorithms*, 1(2):221–229, 1990.

[JW15] Zahra Jafargholi and Daniel Wichs. Tamper detection and continuous non-malleable codes. In *Theory of Cryptography - 12th Theory of Cryptography Conference, TCC 2015, Warsaw, Poland, March 23-25, 2015, Proceedings, Part I*, pages 451–480, 2015.

[Kar71] A.A. Karatsuba. On a certain arithmetic sum. *Soviet Math Dokl., 12, 1172-1174*, 1971.

[Kar91] AA Karatsuba. The distribution of values of dirichlet characters on additive sequences. In *Doklady Acad. Sci. USSR*, volume 319, pages 543–545, 1991.

[KKL88] Jeff Kahn, Gil Kalai, and Nathan Linial. The influence of variables on boolean functions (extended abstract). In *29th Annual Symposium on Foundations of Computer Science, White Plains, New York, USA, 24-26 October 1988*, pages 68–80, 1988.

[KLR09] Y. Kalai, X. Li, and A. Rao. 2-source extractors under computational assumptions and cryptography with defective randomness. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 617–626, 2009.

[KLRZ08] Y. Kalai, X. Li, A. Rao, and D. Zuckerman. Network extractor protocols. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 654–663, 2008.

[KN97] Eyal Kushilevitz and Noam Nisan. *Communication complexity*. Cambridge University Press, 1997.

[Kon03] Sergei Konyagin. A sum-product estimate in fields of prime order. *CoRR*, arXiv:math/0304217, 2003.

[KR09] B. Kanukurthi and L. Reyzin. Key agreement from close secrets over unsecured channels. In *EUROCRYPT 2009, 28th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, 2009.

[KRVZ11] Jesse Kamp, Anup Rao, Salil P. Vadhan, and David Zuckerman. Deterministic extrac-

tors for small-space sources. *Journal of Computer and System Sciences*, 77:191–220, 2011.

[KZ07a] Jesse Kamp and David Zuckerman. Deterministic extractors for bit-fixing sources and exposure-resilient cryptography. *SIAM J. Comput.*, 36(5):1231–1247, 2007.

[KZ07b] Jesse Kamp and David Zuckerman. Deterministic Extractors for Bit-Fixing Sources and Exposure-Resilient Cryptography. *Siam Journal on Computing*, 36:1231–1247, 2007.

[Len90] Thomas Lengauer. *Handbook of Theoretical Computer Science (Vol. A)*. MIT Press, Cambridge, MA, USA, 1990.

[Li11a] Xin Li. Improved constructions of three source extractors. In *Proceedings of the 26th Annual IEEE Conference on Computational Complexity, CCC 2011, San Jose, California, June 8-10, 2011*, pages 126–136, 2011.

[Li11b] Xin Li. A new approach to affine extractors and dispersers. In *Computational Complexity (CCC), 2011 IEEE 26th Annual Conference on*, pages 137–147, June 2011.

[Li12a] Xin Li. Design extractors, non-malleable condensers and privacy amplification. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing*, pages 837–854, 2012.

[Li12b] Xin Li. Non-malleable extractors, two-source extractors and privacy amplification. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science*, pages 688–697, 2012.

[Li13a] Xin Li. Extractors for a constant number of independent sources with polylogarithmic min-entropy. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, pages 100–109, 2013.

[Li13b] Xin Li. New independent source extractors with exponential improvement. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, pages 783–792, 2013.

[Li15a] Xin Li. Improved constructions of two-source extractors. *Electronic Colloquium on Computational Complexity (ECCC)*, 2015.

[Li15b] Xin Li. Improved constructions of two-source extractors. Technical Report TR15-125, ECCC, 2015.

[Li15c] Xin Li. Improved two-source extractors, and affine extractors for polylogarithmic entropy. Technical Report TR15-125, ECCC, 2015.

[Li15d] Xin Li. Non-malleable condensers for arbitrary min-entropy, and almost optimal protocols for privacy amplification. In *12th Theory of Cryptography Conference*, 2015.

[Li15e] Xin Li. Three-source extractors for polylogarithmic min-entropy. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, 2015.

[LL12] Feng-Hao Liu and Anna Lysyanskaya. Tamper and leakage resilience in the split-state model. In *CRYPTO*, pages 517–532, 2012.

[LRVW03] Chi-Jen Lu, Omer Reingold, Salil P. Vadhan, and Avi Wigderson. Extractors: optimal up to constant factors. In *STOC*, pages 602–611, 2003.

[Mau92] Ueli M. Maurer. Conditionally-perfect secrecy and a provably-secure randomized cipher. *Journal of Cryptology*, 5(1):53–66, 1992.

[Mek15] Raghu Meka. Explicit resilient functions matching Ajtai-Linial. *CoRR*, abs/1509.00092, 2015.

[MU02] Elchanan Mossel and Christopher Umans. On the complexity of approximating the {VC} dimension. *Journal of Computer and System Sciences*, 65(4):660 – 671, 2002. Special Issue on Complexity 2001.

[MW97] Ueli Maurer and Stefan Wolf. Privacy amplification secure against active adversaries. In *Advances in Cryptology — CRYPTO '97*, volume 1294, pages 307–321, August 1997.

[NT99] Noam Nisan and Amnon Ta-Shma. Extracting randomness: A survey and new constructions. *Journal of Computer and System Sciences*, 58(1):148 – 173, 1999.

[NZ96] Noam Nisan and David Zuckerman. Randomness is linear in space. *J. Comput. Syst. Sci.*, 52(1):43–52, 1996.

[PR04] P. Pudlak and V. Rodl. Pseudorandom sets and explicit constructions of Ramsey graphs, 2004.

[PR08] Rafael Pass and Alon Rosen. Concurrent nonmalleable commitments. *SIAM J. Comput.*, 37(6):1891–1925, 2008.

[Rao07] Anup Rao. An exposition of Bourgain's 2-source extractor. *Electronic Colloquium on Computational Complexity (ECCC)*, 14(034), 2007.

[Rao09a] Anup Rao. Extractors for a constant number of polynomially small min-entropy independent sources. *SIAM J. Comput.*, 39(1):168–194, 2009.

[Rao09b] Anup Rao. Extractors for low-weight affine sources. In *Proceedings of the 24th Annual IEEE Conference on Computational Complexity*, 2009.

[Raz05] Ran Raz. Extractors with weak random seeds. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, pages 11–20, 2005.

[RRV02] Ran Raz, Omer Reingold, and Salil Vadhan. Extracting all the randomness and reducing the error in Trevisan's extractors. *JCSS*, 65(1):97–128, 2002.

[RSZ02] Alexander Russell, Michael E. Saks, and David Zuckerman. Lower bounds for leader election and collective coin-flipping in the perfect information model. *SIAM J. Comput.*, 31(6):1645–1662, 2002.

[RW03] Renato Renner and Stefan Wolf. Unconditional authenticity and privacy from an arbitrarily weak secret. In *Advances in Cryptology — CRYPTO '03, 23rd Annual International Cryptology Conference, Proceedings*, pages 78–95, 2003.

[RY11] Ran Raz and Amir Yehudayoff. Multilinear formulas, maximal-partition discrepancy and mixed-sources extractors. *Journal of Computer and System Sciences*, 77:167–190, 2011.

[RZ01] A. Russell and D. Zuckerman. Perfect-information leader election in $\log^* n + O(1)$ rounds. *JCSS*, 63:612–626, 2001.

[RZ08] Anup Rao and David Zuckerman. Extractors for three uneven-length sources. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques, 11th International Workshop, APPROX 2008, and 12th International Workshop, RANDOM 2008, Boston, MA, USA, August 25-27, 2008. Proceedings*, pages 557–570, 2008.

[Sak89] Michael Saks. A robust noncryptographic protocol for collective coin flipping. *SIAM Journal on Discrete Mathematics*, 2(2):240–244, 1989.

[Sch80] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *J. ACM*, 27(4):701–717, October 1980.

[Sha06] Ronen Shaltiel. How to get more mileage from randomness extractors. In *21st Annual IEEE Conference on Computational Complexity (CCC 2006), 16-20 July 2006, Prague, Czech Republic*, pages 46–60, 2006.

[Sha08] Ronen Shaltiel. How to get more mileage from randomness extractors. *Random Struct. Algorithms*, 33(2):157–186, 2008.

[Sho90] Victor Shoup. Searching for primitive roots in finite fields. In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing, May 13-17, 1990, Baltimore, Maryland, USA*, pages 546–554, 1990.

[SSZ95] Michael Saks, Aravind Srinivasan, and Shiyu Zhou. Explicit dispersers with polylog degree. In *Proceedings of the Twenty-seventh Annual ACM Symposium on Theory of Computing*, STOC '95, pages 479–488, New York, NY, USA, 1995. ACM.

[SV86] Miklos Santha and Umesh V. Vazirani. Generating quasi-random sequences from semi-random sources. *Journal of Computer and System Sciences*, 33:75–87, 1986.

[Tal14] Avishay Tal. Tight bounds on the Fourier spectrum of AC^0. *Electronic Colloquium on Computational Complexity (ECCC)*, 21:174, 2014.

[Tre01] Luca Trevisan. Extractors and pseudorandom generators. *Journal of the ACM*, pages 860–879, 2001.

[TV00] Luca Trevisan and Salil P. Vadhan. Extracting Randomness from Samplable Distributions. In *IEEE Symposium on Foundations of Computer Science*, pages 32–42, 2000.

[TV06] Terence Tao and Van H. Vu. *Additive Combinatorics*. Cambridge University Press, 2006.

[TZ04] A. Ta-Shma and D. Zuckerman. Extractor codes. *IEEE Transactions on Information Theory*, 50:3015–3025, 2004.

[Uma99] C. Umans. Hardness of approximating sigma;2p minimization problems. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 465–474, 1999.

[Vad04] Salil P. Vadhan. Constructing locally computable extractors and cryptosystems in the bounded-storage model. *J. Cryptology*, 17(1):43–77, 2004.

[Vio14] Emanuele Viola. Extractors for circuit sources. *SIAM J. Comput.*, 43(2):655–672, 2014.

[vN51] J. von Neumann. Various techniques used in connection with random digits. *Applied Math Series*, 12:36–38, 1951. Notes by G.E. Forsythe, National Bureau of Standards. Reprinted in *Von Neumann's Collected Works*, 5:768-770, 1963.

[VV85] U. V. Vazirani and V. V. Vazirani. Random polynomial time is equal to slightly-random polynomial time. In *Foundations of Computer Science, 1985., 26th Annual Symposium on*, pages 417–428, Oct 1985.

[WZ93] Avi Wigderson and David Zuckerman. Expanders that beat the eigenvalue bound: Explicit construction and applications. In *Proceedings of the Twenty-fifth Annual ACM Symposium on Theory of Computing*, STOC '93, pages 245–251, New York, NY, USA, 1993. ACM.

[Yao79] Andrew Chi-Chih Yao. Some complexity questions related to distributive computing. In *ACM Symposium on Theory of Computing*, pages 209–213, 1979.

[Yeh11] Amir Yehudayoff. Affine extractors over prime fields. *Combinatorica*, 31(2):245–256, 2011.

[Zip79] Richard Zippel. Probabilistic algorithms for sparse polynomials. In *Proceedings of the International Symposiumon on Symbolic and Algebraic Computation*, EUROSAM '79, pages 216–226, London, UK, UK, 1979. Springer-Verlag.

[Zuc90] D. Zuckerman. General weak random sources. *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 0:534–543 vol.2, 1990.

[Zuc91] David Zuckerman. *Computing Efficiently Using General Weak Random Sources*. PhD thesis, University of California at Berkeley, 1991.

[Zuc96] D. Zuckerman. Simulating bpp using a general weak random source. *Algorithmica*, 16(4):367–391, 1996.

[Zuc97] David Zuckerman. Randomness-optimal oblivious sampling. *Random Structures and Algorithms*, 11:345–367, 1997.

[Zuc07] David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *Theory of Computing*, pages 103–128, 2007.

Vita

Eshan Chattopadhyay was born in Visakhapatnam, India on 23rd September, 1989, the son of Atrayee Chattopadhyay and Buddhadeb Chattopadhyay. After completing his high school in Hyderabad, he attended Indian Institute of Technology at Kanpur. He received a Bachelor of Technology in Computer Science from IIT Kanpur in 2011. He entered the graduate school at University of Texas at Austin in September 2011.

Permanent Address: USA

This dissertation was typeset with LaTeX 2_ε[3] by the author.

[3] LaTeX 2_ε is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The macros used in formatting this dissertation were written by Dinesh Das, Department of Computer Sciences, The University of Texas at Austin, and extended by Bert Kay, James A. Bednar, and Ayman El-Khashab.